1998

Serially Correlated Wages in a Dynamic, Discrete Choice Model of Teacher Attrition

Todd R. Stinebrickner

ISSN:0318-725X
ISBN:0-7714-2144-3

RESEARCH REPORT 9821

Serially Correlated Wages in a Dynamic, Discrete Choice Model of Teacher Attrition

by

Todd R. Stinebrickner

November 1998

Department of Economics
Social Science Centre
University of Western Ontario
London, Ontario, Canada
N6A 5C2
econref@julian.uwo.ca

Serially Correlated Wages in a Dynamic, Discrete Choice Model of Teacher Attrition

Todd R. Stinebrickner
The University of Western Ontario
Dept. of Economics
Social Sciences Center
London, Ontario
CA N6A 5C2
trstineb@julian.uwo.ca

This paper suggests and implements a method for dealing with the problems which are encountered during the estimation of dynamic, discrete choice models with serially correlated unobservables. The method takes advantage of Gaussian quadrature integral approximation techniques and is based on a new, non–parametric value function approximation algorithm which allows the econometrician to avoid potentially problematic functional form specifications. A desirable property of the approach is that the econometrician has complete control over the factors which ensure that parameter estimates from an approximate solution to a model with serial correlation can be made arbitrarily close to the true parameter estimates. This property potentially allows the econometrician to gauge how close the approximate solution to a model with serially correlated unobservables is to the true solution. The method is illustrated using an example of the occupational choice decisions of certified elementary and high school teachers. Tests of approximation quality suggest that the estimation of dynamic, discrete choice models with serial correlation can often be achieved with little approximation bias, even without incurring large amounts of computational costs.

I would like to thank Steven Stern, Michael Brien, Chris Ferrall, Donna Gilleskie, Jeff Smith, John Geweke, Jonathan Skinner, John Bound, William Johnson, Chris Swann, and seminar participants at the Econometric Society 1997 Summer Meetings, Virginia, Michigan, Arizona, York, Western Ontario, Maryland, Washington, LSU, Georgia St., Hawaii, UNC Greensboro, New Mexico, and Toronto.

Serially Correlated Wages in a Dynamic, Discrete Choice Model of Teacher Attrition

This paper suggests and implements a method for dealing with the problems which are encountered during the estimation of dynamic, discrete choice models with serially correlated unobservables. The method takes advantage of Gaussian quadrature integral approximation techniques and is based on a new, non–parametric value function approximation algorithm which allows the econometrician to avoid potentially problematic functional form specifications. A desirable property of the approach is that the econometrician has complete control over the factors which ensure that parameter estimates from an approximate solution to a model with serial correlation can be made arbitrarily close to the true parameter estimates. This property potentially allows the econometrician to gauge how close the approximate solution to a model with serially correlated unobservables is to the true solution. The method is illustrated using an example of the occupational choice decisions of certified elementary and high school teachers. Tests of approximation quality suggest that the estimation of dynamic, discrete choice models with serial correlation can often be achieved with little approximation bias, even without incurring large amounts of computational costs.

I would like to thank Steven Stern, Michael Brien, Chris Ferrall, Donna Gilleskie, Jeff Smith, John Geweke, Jonathan Skinner, John Bound, William Johnson, Chris Swann, and seminar participants at the Econometric Society 1997 Summer Meetings, Virginia, Michigan, Arizona, York, Western Ontario, Maryland, Washington, LSU, Georgia St., Hawaii, UNC Greensboro, New Mexico, and Toronto.

1. Introduction

In many economic contexts, dynamic discrete choice models represent a theoretically desirable way to describe an individual's decision–making process when the individual faces uncertainty about the future and is forward looking. Unfortunately, the use of certain appealing specifications of such models can often be burdensome or infeasible due to the computational obstacles which are encountered during estimation. The estimation of dynamic, discrete choice models with continuous, serially correlated stochastic components represents one prominent example in which the difficulties which arise during estimation have remained largely unresolved (Rust and Phelan, 1997). Rust (1992) notes that in some cases the lack of effective methods for dealing with serially correlated unobservables may be a "blessing in disguise" because it prompts researchers to be more careful in measuring unobservables and incorporating them explicitly into the economic models. Nonetheless, in cases where agents are likely to realize that the stochastic components of certain state variables are serially correlated, ignoring this serial correlation may lead to a significant misrepresentation of a person's expectations about these variables.

This paper suggests and implements a method for dealing with this problem that is both computationally feasible and generally applicable. The method takes advantage of Gaussian quadrature integral approximation techniques and is based on a new, non–parametric value function approximation algorithm which allows the econometrician to avoid potentially problematic functional form specifications. A desirable property of the approach is that the econometrician has complete control over the factors which ensure that parameter estimates from an approximate solution to a model with serial correlation can be made arbitrarily close to the

true parameter estimates. This property will often allow the researcher to gauge how well an approximation is performing in his specific application.

The computational burdens which can be encountered during the estimation of dynamic, discrete choice models are well known. As with static, multinomial discrete choice models, the computation of choice probabilities in dynamic, discrete choice models requires the potentially costly evaluation of a multi–dimensional integral. An additional computational burden arises when a person is assumed to consider both current period utility and discounted expected future utility because it is no longer trivial to compute the utility, or value, associated with each available option at some time t. Bellman's functional equation reveals that the expected future utility component of a time t value function can be written as the expected maximum of the values associated with the options that the person will have in time t+1 conditional on his time t decision. When a finite–horizon setting is appropriate, value functions are nonstationary and the traditional solution technique for value functions involves backwards recursion.[1] Two inter–related sources of computational burden arise during the backwards recursion solution process. First, because the deterministic portions of the time t+1 value functions serve as the inputs into the expected future utility function, it is necessary to compute the deterministic portions of the value functions at time t+1 for **all** possible combinations of the state variables that could arise at t+1 before the set of necessary value functions can be solved at time t. This is what Bellman (1957) called the "curse of dimensionality." Secondly, once the deterministic portions of the necessary time t+1 value functions have been solved, computing the expected future utility component of each time t value function requires the evaluation of a

[1]When an infinite horizon setting is appropriate, solutions to value functions represent a fixed point in the space of value functions.

multi–dimensional integral with respect to the model's time t+1 stochastic components whose realizations are unknown to the agent at time t.

The presence of a continuous, serially correlated unobservable accentuates these problems. An immediate implication is that the serially correlated unobservable must be treated as a state variable because its current value contains information about future utility flows. However, because a closed form solution does not typically exist for value functions, it is not possible to treat a state variable as truly continuous using the traditional solution method; the resulting infinite number of state variable combinations causes a direct conflict with the backwards recursion solution method. One possible way to deal with this problem is to discretize the continuous unobservable. Since the stochastic process associated with the underlying continuous variable can be used to generate the inter–period transition probability matrix associated with the possible values of the discretized variable, this approach does allow the use of convenient time–series specifications for the serially correlated components.[2] For example, Hubbard et al. (1995) allow two state variables to follow first–order autoregressive processes around a deterministic trend and discretizes the deviations from the trend into nine nodes. In this case, inter–period transition probabilities between the discrete states follow first–order Markov processes.

However, there are several drawbacks of the discretization approach which can be seen by considering a model in which only one state variable exists, it is serially correlated, and it is discretized into q possible values. First, if the discretized variable is treated as a continuous variable in the specification of the model then desirable variation in the data is lost and the

[2]For example, MaCurdy (1982) discusses the virtues of using time series processes to model the error structure of earnings in longitudinal data.

included variable is measured with error.[3] Secondly, even after the deterministic portions of the necessary time t+1 value functions have been solved, the computation time necessary to evaluate the expected future utility component of the time t value function associated with some value of the discretized state variable will increase by a factor of q relative to an otherwise identical model which does not include a serially correlated state unobservable. This occurs because the expected future utility component becomes a weighted average of the expected future utilities that a person would receive conditional on each of the q possible values of the discretized unobservable that could arise at time t+1, where the weights are the transition probabilities between the value of the discretized unobservable at time t and the q possible values at time t+1. Further, the computational burden increases exponentially with the number of serially correlated state variables that are included in the model; if r discretized variables exist, the number of elements in the weighted average is q^r. While the computational burden associated with this (potentially multi–dimensional) sum can be reduced by discretizing the continuous variables into a smaller number of nodes (a smaller number q), this reduction will further increase the amount of measurement error in the included variable.[4] Finally, the use of a discretized unobservable may increase the difficulty of estimating the continuous stochastic process of interest jointly with the other structural parameters because the likelihood function (or moment conditions) may lack

[3] If the discretized variable enters the model as a series of indicator variables, a certain amount of desirable variation in the data will be lost. Further, including interaction terms of the variable of interest and other variables will typically lead to large increases in the number of parameters to be estimated.

[4] If the variable of interest enters the model as a series of indicator variables, the effect of reducing q will be to further reduce the variation in the variable.

continuous derivatives with respect to the parameters of the stochastic process.[5] By explicitly modeling individual behavior, dynamic programming models potentially have the ability to control for certain selection effects that exist in the stochastic processes of interest.[6] This benefit is lost if the stochastic process cannot be estimated jointly with the other structural parameters.

Besides the work of Hubbard et al., serial correlation has been accommodated to some degree in only a small number of special cases of dynamic, discrete choice models. For example, Berkovec and Stern (1991) use an error structure similar to Miller (1984) to estimate a model in which future realizations of serially correlated error terms are unknown by the econometrician but are entirely deterministic from the standpoint of the agent in the model.[7] While this type of specification avoids many of the computational problems discussed above, it does not deal directly with the problem of allowing a serially correlated variable to influence future uncertainty from the individual standpoint.

---

[5]For example, in the empirical example below, the serially correlated, continuous unobservable in the wage equation is not observed for an individual in all periods. Thus, in order to compute the likelihood contribution for a person with missing wages, it is necessary to repeatedly simulate wage error values for all of the person's missing wage years from their joint distribution conditional on all of the observed wage error terms. When the error terms are treated as continuous, a small change in one of the parameters of the wage equation will lead to only a small change in the sequence of simulated wage errors. However, problems arise if the continuous wage error simulations must be mapped to discrete values because small change in the simulated wage errors may be enough to push some of the continuous variables across thresholds which determine the value of the discrete variable. The resulting discontinuous jump in the discrete variable will lead to a discontinuous change in the likelihood contribution for the person. Thus, Newton–Raphson optimization or other derivative based optimization routines are likely to perform poorly.

[6]For example, by explicitly modelling the decision of whether or not to work, a dynamic programming model of labor supply may be well–suited to deal with the selection effect in wages that exists when wages are only observed for those who work.

[7] Among others, Gilleskie (1998) takes a similar approaches in order to allow unobservables to follow a limited form of serial correlation. Other relevant microeconomic applications which incorporate some type of serial correlation include Pakes (1986) and Christensen (1990). The life-cycle models of Palumbo (1991) and Engen (1992) also allow limited types of serial correlation. However, unlike dynamic, discrete choice models, these models are characterized by continuous decision variables which implies that optimal decisions of the individuals can be described by Euler conditions. Relevant macroeconomic applications includes work by Giovannini and Labadie (1991) and Hodrick et al. (1991).

The recent approximation method of Keane and Wolpin (1994) represents a very important contribution from the standpoint of allowing the estimation of more general forms of dynamic programming models. At each point in time t, their method requires that the expected future utility components of value functions be computed for only a subset of the possible time t+1 state points. These computed values are then used to fit a single interpolating regression which can provide a fitted value for the expected future utility associated with any other time t+1 state point.[8] Under many general model specifications, this method provides substantial time savings because the integration necessary to evaluate the expected future utility term will be computationally burdensome whereas computing a fitted value from the interpolating function is not.

Keane and Wolpin (1994) discusses the possibility of using this general method to incorporate serial correlation in dynamic programming models which are being solved by backwards recursion.[9] As mentioned earlier, the most obvious problem which is encountered in estimating a model which includes serial correlation is that the continuous unobservable implies

---

[8]Time savings may not occur when stochastic components are assumed to be iid extreme value. In this case, the most desirable Keane and Wolpin interpolating function takes as long to evaluate as the closed form solution which exists for the Emax.

[9]Erdem and Keane (1996) uses this method in a model of dynamic brand choice to allow consumers to update their expectations of brand attributes in a Bayesian manner.

Rather than solving the exact dynamic programming problem, Stock and Wise (1990) use an alternative formulation for the future utility component of value functions which is easy to evaluate and allows their model to incorporate important future uncertainty involving serially correlated error terms in a forward looking model of retirement behavior. Lumsdaine, Stock and Wise (1991) show that in their particular example, the bias of the approximation is small. However, Stern (1997a) finds that there are many predictable cases where the Stock–Wise approximation will perform quite poorly. Hotz and Miller (1993) avoid the backwards recursion solution process altogether by showing that the value associated with a particular option at some time period can be expressed as the sum, over all periods and all future states, of the expected payoff in the state times the probability of the state occurring. Hotz, Miller, Sanders, and Smith (1994) extend this idea by showing that expected rewards need to be considered for only the future states associated with a path of simulated future choices. Although the issue is not explicitly discussed in their work, it appears that the simulation nature of this latter work would theoretically allow the estimation of models with serial correlation.

an infinite number of possible state points. As will be discussed in more detail, the approximation of the expected future utility term by a Gaussian quadrature approximation or by simulation essentially discretizes this continuous variable by specifying a finite number of values of the serially correlated unobservable at which value functions will need to be solved in each period. However, it important to note that this number will typically be extremely large for the majority of the periods in a model. For example, in the empirical example of this paper, the number of necessary error values exceeds one million in many periods (and approaches one billion in some) even under one of the most beneficial specifications. The Keane and Wolpin approximation method represents a way to deal with this situation. Their single interpolating function can be estimated by using the computed expected future utilities associated with a feasible number of possible realizations of the continuous error term at time t+1, and this interpolating function can then be used to interpolate the expected future utility associated with any other particular realization of the continuous error term that is needed at time t+1.

The Keane and Wolpin method involves using actual, rather than interpolated, points whenever possible. The convergence property in their model is based on the fact that as the number of points which are interpolated becomes small, all included points are actually computed by backwards recursion and no approximation error exists. However, this convergence property is not particularly relevant when a serially correlated error term is present because the extremely large number of possible values for the error term implies that essentially all expected future utility components must be interpolated using the single interpolating function. In other words, increasing the number of state points at which the expected future utility term is actually computed to any imaginable upper limit of feasibility may increase the estimated precision of the

parameters in the interpolating function regression, but will lead to only a negligible decrease in the number of expected maximum terms which need to be interpolated. Therefore, poor choices of the interpolating function could lead to poor approximations for essentially all of the included value functions in the model, and, ultimately to biased parameter estimates, **regardless** of the number of points (up to any feasible number) for which value functions are actually solved and used to estimate the interpolating regression. One direct implication is that it is impossible to determine how far the approximate solution is from the true solution in a model which includes serial correlation. What is clear is that the success of the method relies directly on the ability of the researcher to specify a single interpolating function which can accurately predict expected future utility values over the entire set of possible state points. Therefore, it is not surprising that, in general testing which does not involve models with serial correlation, Keane and Wolpin found that approximation quality can vary substantially depending on the specification of the interpolating function. Unfortunately, when serial correlation of a continuous variable is present, the "preferred" and very successful specification of their interpolating function, in which time t+1 value functions directly enter the interpolating function, is not usable.[10]

The quality of an approximation method may vary significantly depending on the particular application. Thus, the primary goal of this paper is to design a method which is both computationally feasible and allows the econometrician to get a sense of how close the approximate solution of a model with serial correlation is to the true solution in his application. In order to accomplish this, the method in this paper gives the econometrican control over the factors which ensure that the approximate solution to a model with one or more serially

[10]As previously mentionned, it is infeasible to solve time t+1 value functions for the extremely large number of possible values of the continuous variable.

correlated error terms can be made arbitrarily close to the true solution. Of utmost importance, it must be the case that the time t+1 value functions which serve as inputs to the expected future utility function can be made arbitrarily close to their true values. To achieve this, a non–parametric approach is taken for approximating value functions. In the spirit of the Keane and Wolpin (1994) approach, value functions are actually solved for only some subset of the state points in the model. However, the interpolating function in this paper is based on a non–parametric approach in which zero weight is assigned to computed points which are not "close" to the point for which value functions are being approximated. By utilizing the computed values of points which immediately "surround" the point for which value functions are being approximated, the approximation method does not rely heavily on the ability of the researcher to correctly specify the functional form of the regression function over all possible values of the state variables, and the interpolation method in this paper has the desirable asymptotic property that an approximated value function necessarily approaches its true value as the "surrounding" state points for which value functions are actually computed by backwards recursion become "close" to the state point for which the value function is being approximated. As will be discussed, the use of the non–parametric approach creates some new issues which must be addressed during the solution and estimation process. For example, in order to guarantee that all state points for which value functions may need to be interpolated are "surrounded" by state points for which value functions are actually computed, it is necessary to keep track of the range of possible values for any state variable for which only a subset of the possible realizations will be computed. For the continuous, serially correlated state variables, it will be shown that the range of values at some time t depends on both the parameters of the model and the range

of possible values that could arise at time t-1. This implies that, for each guess of model parameters in the Newton–Raphson optimization algorithm, before value functions are solved by backwards recursion, the possible range of error values must be solved in each period by a forward recursion process starting in the first period. A second condition that must be satisfied in order for the approximated value function at time t to be close to its true value is that the numerical methods used to approximate the integral associated with the expected future utility term must provide an approximation which is arbitrarily close to the true value of the integral. This is an immediate property of both Gaussian quadrature integral approximation methods (as the number of quadrature points becomes large) and simulation methods (as the number of simulations becomes large).

The solution methods in this paper are explained and tested using an empirical example which examines the labor supply decisions of elementary and high school teachers.[11] Concern about this issue has been heightened by the possibility of a change in teacher demand as children of the baby boom generation continue to enter schools. For example, a recent study by the United States Department of Education projects that the number of public and private classroom teachers will increase by 350,000 between 1995 and 2007.[12] However, ensuring that schools have an ample supply of teachers is not the only reason that this issue is important. Teacher attrition also can drain schools of their most academically gifted teachers because of the rigidity of the

[11]Previous work in the area of teacher decisions that primarily takes advantage of duration models includes Eberts (1987), Murnane and Olsen (1989, 1990), Murnane et al. (1989), Dolton and van der Klaauw (1995), Mont and Rees (1996), Gritz and Theobald (1996), Theobald and Gritz (1996), Brewer (1996), Stinebrickner (1996b,1998). van der Klaauw (1996a) focuses on incorporating information about expectations in dynamic programming models.

[12]This represents a twelve percent increase. See "Projections of Education Statistics to 2007," U.S. Department of Education, NCES 97-382, and the associated special report "A Back to School Special Report on the Baby Boom Echo: Here Come the Teenagers."

wage structure in most public schools. Salary is a function purely of years of teaching experience and post-bachelor education levels and does not allow wage premiums for academically gifted teachers who on average have better non-teaching alternatives than other teachers. Thus, alternative wage structures may increase the labor supply of academically gifted teachers relative to other teachers. Dynamic, discrete choice models are potentially well–suited for performing this type of policy analysis. While the model used in this paper captures many of the important aspects of the teacher decisions, it is important to note that it does not pretend to be perfect. The paper attempts to be very upfront about some of the potential limitations of the model. Stinebrickner (1996a) explores the importance of some of these limitations.

What is important is that the model is useful for exploring the proposed solution to the traditional problems of serial correlation. In particular, the model allows the deviation of wages from a deterministic component to be serially correlated over time within a particular teaching job. The strong effect that is found for serial correlation in this error term is useful because it accents some of the problems discussed above and, therefore, tests the usefulness of the methods in a practical setting. For example, as will be discussed, the range of possible error terms is substantially larger in all periods when strong serial correlation exists.

From a theoretical standpoint, convergence can be established in this paper by both allowing the number of quadrature points (or, alternatively, the number of simulation draws) to become large and by forcing the "surrounding" state points for which value functions are actually solved by backwards recursion to become close to the points for which value functions are interpolated. However, from a practical standpoint, the usefulness of the method depends on how well the approximation performs for feasible choices of the number of quadrature points and the

distance between the state points which are actually computed by backwards recursion. The findings below suggest that the approximate solution remains close to the "true" solution for very feasible choices of these factors. The quality of an approximation method is likely to depend on a particular application. Nonetheless, there appears to be significant practical promise for the methods in this paper.

The remainder of the paper proceeds as follows. Section (2) provides a description of the sample that is used in the paper and examines the decisions of different types of teachers in the data. Section (3) describes the model used in the paper. In section (4), the numerical techniques necessary to deal with the effect that serial correlation has on the solution of the value functions are described, and the new value function approximation methods are introduced. Section (5) describes the methods which are needed to estimate the model. The examination of approximation quality, the estimates of the model of interest, and several policy simulations are presented in section (6). Section (7) summarizes the findings of the paper.

2. Data

2.1 Description of Entire Sample

The National Longitudinal Study of the Class of 1972 is used to estimate the model. This data is sufficiently rich to allow the examination of the crucial elements of this issue. The first wave of this survey, which was completed in 1972, includes interviews with 22,652 students who were expected to graduate from high school in that year. Included in the first wave is information on aptitude tests such as the Scholastic Aptitude Test (SAT). Follow up surveys were taken in 1973, 1974, 1976, 1979, and 1986. Thus, for each person, the survey contains

detailed information about work experience, education, marriage, and fertility for approximately fourteen years after the person graduated from high school. Since survey waves were not collected in every year, some of the survey waves ask the individual retrospective questions which cover several years of the individual's life. One consequence which must be considered during estimation is that the individual is not asked about wages for every year in which he works. The sample used in this paper consists of 451 individuals who became certified to teach at some point between 1975 and 1985. Since this paper involves the career choices of individuals only after they become certified, which usually requires a minimum of four years of training, the data contain between one and eleven years of work histories and personal information for each person.[13] The issues related to the selection of the sample are discussed in Stinebrickner (1996a). Descriptive statistics for the sample are shown in table (1).

In each of the observed data years, whether the person chooses a teaching job, a non–teaching job, or not to work is observed.[14] Of the aggregated 4,041 years of data, 47.8 percent of the years are spent teaching, 32.1 percent of the years are spent working in non–teaching jobs, and the other 20.2 percent of the years are spent not working. The percentage of individuals in a particular group who choose to teach in at least one year will be referred to as the *participation rate* of the group.[15] For the overall sample, the participation rate is 75.6 percent.

---

[13]A small number of people finished college in three years. For these people, 11 years of data are observed. For most people, ten or less years are observed.

[14]Whether the individual has changed jobs is also observed.

[15]Participation rates can be somewhat misleading in this application because they include information on individuals who have been certified for varying lengths of time.

The teaching duration for those who decide to teach can also be examined. In particular, figure (1) shows a Kaplan-Meier survivor function associated with the first spell in teaching for the 341 individuals who enter teaching at some point. In this work, the end of the first spell in teaching occurs when a person chooses a non-teaching job or not to work at all; changing teaching jobs does not represent the end of a spell. The survivor function evaluated at a duration of t years is the probability that an individual will teach more than t years before quitting. Thus, in the first teaching spell, the probability that a teacher will teach more than three years before leaving teaching is .611, and the probability that the teacher will stay in teaching more than four periods is .504.[16]

2.2 Behavior of Different Types of Teachers

The quality of teachers has been shown to be a very important input to student learning (Bishop, 1996). Unfortunately, the data do not provide a direct measure of teaching effectiveness. However, test scores assessing the teacher's academic ability are observed and have been shown to be a teacher characteristic which increases student learning (Hanushek, 1971; Strauss and Sawyer, 1986; Ferguson, 1990; Ehrenberg and Brewer 1993; Monk, 1992).[17] For the

[16]Since the data is yearly, the wording "more than 4 years" could be replaced with "5 or more years." This number appears to be roughly consistent with those found using teacher-specific data. For example, Murnane and Olsen (1989) find that approximately .59 of English teachers complete five or more years of teaching. They find that this number is higher for elementary teachers, math teachers, and social studies teachers and lower for biology and chemistry/physics teachers. The numbers in this figure may be slightly lower because it was constructed assuming an individual exits teaching if he or she chooses not to teach in a particular year. About .18 of exiting teachers return after being out one year.

[17] For example, Strauss and Sawyer (1986) find that having teachers with higher scores on the National Teaching Examination (NTE), which Ayers and Qualls (1979) find to be positively correlated with SAT scores, has a substantial effect on whether or not students fail to demonstrate independently measured reading and mathematics skills. Murnane et al. (1991) does not conclude that the research link is quite as convincing as Bishop (1996). Nonetheless, the trend by many state legislatures to increase the stringency of minimum standardized test scores for new teachers seems to indicate a general belief among policy-makers that recruiting high ability teachers should be a key concern of schools.

remainder of the paper the term "ability" will be used to refer to the academic ability of the teacher.

The measure of ability that is used in this analysis is the individual's score on the math SAT score.[18] Using SAT scores as the grouping criteria, the occupational trends of different ability groups are examined. Figure (2) shows the percentage of aggregate years spent in teaching jobs, non-teaching jobs, and the home option for five SAT groups: teachers in the lowest third of the sample in terms of SAT scores, teachers in the middle third, teachers in the highest third, teachers in the highest fifth, and teachers in the highest tenth. Notice that the percentage of periods spent in teaching decreases monotonically across groups as the ability level increases and, therefore, shows that high ability teachers do choose teaching less frequently than other teachers under the current rigid wage structure. The sample includes not only people who are actually decide to teach, but also people who are certified to teach but choose alternative career paths. Stinebrickner (1996a) shows that higher ability teachers are less likely to actually enter teaching, and those that do enter teaching have shorter average spell lengths. However, this paper does not account for the reality that the decision to become certified is endogenous.[19]

---

[18]It is likely that retaining and recruiting teachers with high math SAT scores will be an important priority given student deficiencies on standardized tests of basic math and science. In general, the retention of teachers with high math SAT scores may be more difficult than the retention of teachers with high verbal SAT scores because the math SAT score is a strong predictor of opportunity costs. For example, Murnane et al. (1995) finds that males graduating from high school in 1972 with strong basic math skills have significantly higher hourly earnings at age 24 than males graduating in that year with average math skills. Using data of this paper, simple wage regressions indicate that teachers with high math SAT scores do receive higher non–teaching wages than other teachers. However, this non–teaching wage premium is not present for individuals with high verbal SAT scores. Using combined math and verbal scores would tend to obscure some of the problems/solutions associated with retaining high math SAT teachers.

[19]Allowing an endogenous teacher certification decision would create a much more complex choice set for the individual in the model. Another possibility would be to try to include a simple reduced–form certification equation. However, credibly estimating the relationship between the reduced form equation and the structural equations would seem to be a difficult task. Thus, I chose to examine the choices of individuals after the certification decision.

Therefore, the effect that changing teaching wage structures has on the certification decision will not be examined.

3. The Model

3.1 An Individual's Alternatives

Each teacher's employment decision is modelled as a finite horizon, discrete time, dynamic programming problem. Let time period 0 be the year in which the individual becomes certified to teach and T^* be the end of the decision making horizon. Then, in each period $t=1,2,...,T^*$, the individual makes an employment decision that maximizes the sum of current utility and discounted expected future utility. At each time t, the individual has the option of accepting a teaching job, E, accepting a non-teaching job, N, or choosing not to work, H. It is assumed that the individual receives one new teaching offer and one new non–teaching offer per period. In addition, if the individual chooses to teach in period t-1, then in period t he also receives a teaching offer associated with returning to the old teaching job. By differentiating between new and old teaching offers, the model explicitly allows individuals to search for new, possibly more lucrative, teaching jobs while currently teaching. The data suggest that this search process plays an important role in the labor decisions of certified teachers. In the data, forty-four percent of all individuals who teach in at least one year hold multiple teaching jobs during the sample period and thirty-seven percent of the aggregate teaching years during the sample period are spent in teaching jobs other than a person's initial teaching job.[20] Notationally, the model will distinguish between different teaching jobs by specifying the starting dates of the teaching

[20]This number was computed assuming that individuals who leave teaching and return at a later date start a new job at this point. Nonetheless, this assumption is not driving this result. The number of individuals who report holding multiple jobs during their first teaching spell is significant.

jobs; Eb will represent a job offer associated with a teaching job which was started at time b. The model does not make a distinction between new and old non–teaching jobs. Let $\bar{D}^j_t$ j=Eb, N, and H, represent the set of choices available to the person at time t given his choice of j at time t-1, and let d^j_t be the number of elements in $\bar{D}^j_t$.[21]

3.2 Current Period Utility

As in Berkovec and Stern (1991), the individual is assumed to consider both the wage and non-wage utility that he would derive from each option in the current period. Associated with each teaching job offer is a wage offer which depends in part on the years of teaching experience accumulated at time t, Y_t, a vector of other observable personal characteristics (including the post-bachelor education level), X, and the time period, t.[22] Consider the time t teaching wage, W^{Eb}, associated with a job which began at time b:

$$W^{Eb}(t,X,Y_t,v^{Eb}_t)=\bar{W}^E(t,X,Y_t)+v^{Eb}_t. \tag{3.1}$$

The deterministic portion, $\bar{W}^E$, measures the average teaching wage that an individual with a particular set of observable characteristics could expect to receive at time t. Years of teaching experience and the post–secondary education level of the individual at time t are important determinants of teaching wages because wage structures are typically rigid functions which take into account only these two things. Other personal characteristics, such as an observable measure

[21] $\bar{D}^{Eb}_t$={Eb,Et,N,H} which implies d^{Eb}_t=4. $\bar{D}^i_t$={Et,N,H} and d^i_t=3 for i=N and H.

[22] Although some of the elements of X are time-varying, for notational simplicity the time subscript is not included on X. It is the total years of teaching experience which enters the wage structure in most public schools. Thus, the relevant tenure variable is Y_t rather than the number of years of experience at a particular job, t-b.

of ability, are also elements of X and enter the wage equation to allow for the possibility that teachers with desirable characteristics may be able to obtain jobs in higher paying school districts.

The error term in the teaching wage equation, ν^{Eb}_t, measures the difference between the wage that a person receives in a particular teaching job at time t and the average wage that the person could expect to receive in teaching at time t conditional on his observable characteristics. It is this error term which is allowed to be serially correlated in the model. There are several reasons to think that serial correlation might be important in this error term. One reason for this is simply that some schools are generally higher paying than others.[23] However, there are also reasons to think that a teacher's expectations about future wages are updated in response to actual yearly decisions that a school district makes with respect to wages. The period covered by the data was one of fluctuating real wages for teachers. Real wages typically fell significantly during the 1970's, as nominal wage increases did not keep pace with inflation, before beginning to rebound during the early 1980's (see Cohn, 1996). Thus, it is likely that an individual would gain information about future wages by observing how his particular school responded to changing inflation rates. In an attempt to capture this sort of individual learning about future wages, the stochastic process governing ν^{Eb}_t is specified as an autoregressive, AR(1), process.[24] This specification is detailed in section (3.4).

---

[23]If differences in the level of pay between schools was the only source of inter–period correlation, a job specific heterogeneity term could simply be included in the wage equation. Even if this was the only effect we were interested in, including person specific heterogeneity would not seem to be sufficient given the evidence of job mobility in the data. Including continuous heterogeneity would be difficult if the heterogeneity were job specific rather than person specific. A more promising approach in this case would be to use discrete heterogeneity techniques such as those employed by Heckman and Singer (1984), Cameron and Heckman (1998), and Mroz and Guilkey (1992).

[24]Although this is a strong assumption for the wage process, it does represent a significant relaxation of the standard independence assumption. Although it is not done here, one might want to include both job specific heterogeneity and serial correlation.

The non-teaching wage equation has a form which is similar to the teaching wage equation:

$$W^N(t,X,v^N_t)=\bar{W}^N(t,X)+v^N_t. \tag{3.2}$$

Note that the years of non–teaching experience is not included in the model. Further, as will be discussed in section (3.1), the model also makes the standard assumption that v^N_t is serially uncorrelated. It is important to note that these specification decisions were not made on theoretical grounds. Instead, given the potential difficulty that would be encountered in specifying the model with a serially correlated teaching wage error, a decision was made to attempt to provide a reasonable measure of teachers' outside wage alternatives which is as simple as possible.[25]

The person receives a wage of zero if he chooses the home option so that $\bar{W}^H=0$, $v^H_t=0$, and

$$W^H=0. \tag{3.3}$$

The non-pecuniary utility associated with each option is the sum of three components:

$$Q^j(t,X,Y_t,\varepsilon^j_t) = \bar{Q}^j(t,X,Y_t) + \mu^j + \varepsilon^j_t \quad \text{for } j = Eb, N, \text{ and } H. \tag{3.4}$$

$\bar{Q}^j$ measures the average effect that observable characteristics, such as years of teaching experience and number of children, have on the non-wage enjoyment that a person receives from

[25]Stinebrickner (1996a) examines the impact of not including years of non–teaching experience in the model. The issues related to allowing the non–teaching wage to be serially correlated are identical to those for the teaching wage equation and are discussed in section (4.3). Certainly, there is no guarantee that ignoring serial correlation in non–teaching alternatives is less important than ignoring serial correlation in teaching alternatives. However, wage specifications in previous work do not include the type of serial correlation proposed here, and, in many previous models, wages are estimated separately and only a measure of an individual's average wage enters the model.

option j. The term μ^j is an unobserved heterogeneity component which measures the individual specific enjoyment of the non-wage benefits of option j. It allows individuals with identical observable characteristics to receive different average levels of enjoyment from a particular option.[26] The error term ε^j_t is a random shock to the non-wage utility of option j. One interpretation of ε^j_t is that it captures information about state variables that are unobserved by the econometrician. For example, ε^{Eb}_t may include specific information about the grade level, class size, and student ability level associated with a particular teaching offer at time t or may represent other factors which influence an individual's attitude towards teaching. Under this interpretation, it is reasonable to assume that when making decisions the individual knows the current values of ε^j_t but knows only the distributions of future values.

The total current period utility associated with a particular option is the sum of the wage and non-wage utility:

$$U^j(t,X,Y_t,C^j_t)=W^j(\bullet) + Q^j(\bullet) \quad \text{for j = Eb, N, and H} \tag{3.5}$$

where $C^j_t = \{\nu^j_t,\varepsilon^j_t\}$ represents the set of wage and non-wage errors that are relevant in determining the total current period utility of option j. Since wages enter the total current period utility function directly, the model makes the assumption of no saving by individuals.[27]

---

[26]Note that the reason for not including μ^j explicitly as an argument of $Q^j(\bullet)$ is strictly one of notational simplicity. This is also true of the total current period utility functions and the value functions which will be described in the next several paragraphs.

[27]This is a standard assumption in these models. See for example, Berkovec and Stern (1991), Swann (1997), Rust and Phelan (1997), and van der Klaauw (1996b).

3.3 Discounted Expected Utility - Value Functions

The dynamic nature of the model implies that an individual considers the discounted expected utility, or value, of all available options when making decisions. This value can be written as the sum of the current period utility and the discounted expected future utility. Bellman's functional equation allows value functions to be written in a recursive manner by noting that the expected future utility can be written as a function of the values of the alternatives that the individual will have in the *next* period conditional on his current period choice. Conditional on choosing j ε (Eb,N,H) at time t, the person does not know the exact future utility that he will receive (i.e., he can only compute the *expected* future utility) because he does not know the realization of a set of relevant error terms, S^j_{t+1}, that determines the values of the various t+1 alternatives. S^j_{t+1} depends on the time t choice, j, through the effect that j has on the types of alternatives that are available to the person at time t+1 and through the effect that j has on the state variables in the model. Let V^{Eb}, V^N, and V^H represent the value of accepting a teaching job that was started at time b, a non-teaching job, and the home option respectively.[28] Then,

$$(3.6) \qquad V^j(t,X,Y_t,C^j_t) = U^j(t,X,Y_t,C^j_t) + \beta E_t Z^j(t,X,Y_{t+1},S^j_{t+1}),\ j=Eb,\ N,\ \text{and}\ H$$

where the term $\beta E_t Z^j(t,\bullet)$ represents the discounted (where β is the discount factor) expected value at time t of the highest valued option the individual will have in period t+1 conditional on his choice of option j in the current period. If j is chosen in the current period, then the person must consider the value of the d^j_{t+1} possible choices that he will have in the next period, and each

[28] If b=t then the choice is to accept the new teaching offer.

of these values will reflect the possible changes in the total years of teaching experience that have been accumulated:

$$(3.7)\quad Z^j(t,X,Y_{t+1},S^j_{t+1}) = \max^j \{V^k(t+1,X,Y_{t+1},C^k_{t+1}): k\varepsilon\bar{D}^j_{t+1}\} \quad j=Eb,\ N,\ \text{and}\ H$$

where $\max^j$ represents the maximum function evaluated over the alternatives that will be available in time t+1 conditional on choosing option j at time t. The elements of S^j_{t+1} are the error terms that enter the value functions in the maximum function, $\max^j$:[29]

$$(3.8)\qquad S^j_{t+1} = \bigcup_{k\in\bar{D}^j_{t+1}} C^k_{t+1} = \{\bar{v}^j_{t+1},\bar{\varepsilon}^j_{t+1}\}$$

where $\bar{v}^j_{t+1}$ and $\bar{\varepsilon}^j_{t+1}$ are the sets of all wage and all non–wage errors which appear in the value functions at time t+1 given the choice of j at time t.[30]

The calculation of the expected maximum term, $EZ^j(\bullet)$ for j= Eb, N, and H, involves integrating the maximum function over the joint density, g, of the unknown error terms in S^j_{t+1}:[31]

$$(3.9)\qquad EZ^j(t,X,Y_t,S^j_{t+1}) = \int \max^j(\bullet)\, g(S^j_{t+1}\,|C^j_t)\, dS^j_{t+1} \quad j=Eb,\ N,\ \text{and}\ H.$$

Note that for j=Eb this is a seven dimensional integral because S^{Eb}_{t+1} contains three wage errors and four non-wage errors, and for j=N or H this is a five dimensional integral because S^N_{t+1} and S^H_{t+1} contain two wage errors and three non-wage errors.

---

[29] Recall that $\bar{D}^j_t$ is the set of choices available to the person at time t+1 conditional on choice j in time t.

[30] $\bar{v}^{Eb}_{t+1}=(v^{Eb}_{t+1},\ v^{Et+1}_{t+1},\ v^N_{t+1})$, $\bar{v}^i_{t+1}=\{v^{Et+1}_{t+1},v^N_{t+1}\}$ i=N, and H, $\bar{\varepsilon}^{Eb}_{t+1}=(\varepsilon^{Eb}_{t+1},\varepsilon^{Et+1}_{t+1},\varepsilon^N_{t+1},\varepsilon^H_{t+1})$ and $\bar{\varepsilon}^i_{t+1}=\{\varepsilon^{Et+1}_{t+1},\varepsilon^N_{t+1},\varepsilon^H_{t+1}\}$ for i=N,H.

[31] g will be used generically throughout the paper to represent a density function of its arguments.

3.4 Distributional Assumptions

In order to give more detail about equation (3.9), it is necessary to specify the distributions of the errors in S^j_{t+1}. Notice that the argument of the density function, g, is written in a conditional form which suggests the possibility of inter–period correlations between the error terms in C^j_t and S^j_{t+1}. However, the actual implementation of this type of correlation is perhaps the most well-known obstacle in dynamic, discrete choice models. Much of the difficulty is due to the traditional backwards recursion solution technique which, at each time period, requires the calculation of each value function for all possible combinations of its arguments that could arise. This solution technique implies state variables (which can be defined as endogenous or non–deterministic variables whose present value affects future utility flows) have a costly effect on the amount of computer time necessary to solve the model using backwards recursion. Consider the arguments of the value function in equation (3.6). Y_t is considered endogenous and fits the description of a state variable in the model. The endogeneity assumption implies that the individual does not know his future levels of teaching experience and, therefore, requires value functions at each time period to be solved for *all* possible levels of teaching experience that the person could consider accumulating as of that time period. In contrast, the vector X is considered exogenous and, therefore, does not fit the description of a state variable. The exogeneity assumption implies that at any time period the individual knows the elements of X for all future periods. This implies that value functions need to be solved for *only* the observed value of X at each time period.

The nature of the computational burden of the other value function argument, C^j_t, depends largely on the assumptions that are made about the inter-period correlation structure of the errors.

Under the standard assumption that the error terms in the model are uncorrelated across time, C^j_t does not fit the definition of a state variable. This is the case because C^j_t does not influence the set of t+1 errors, S^j_{t+1}, and, therefore, enters the value function only through its linear relationship to the current period utility, U^j. From a computational standpoint, this strictly linear influence is advantageous because it implies that knowledge of the value function V^j for some value of C^j_t is sufficient to determine V^j for any other possible value of C^j_t. However, under more desirable correlation structures in which current error terms, C^j_t, are correlated with future error terms, S^j_{t+1}, this is no longer the case. Now, the non-linear effect of C^j_t (through the effect on future errors S^j_{t+1}) on the value function implies that C^j_t must be considered a state variable and that V^j must be solved separately for each possible value of C^j_t that could arise. This can lead to significant increases in computational time, especially if C^j_t is a multivariate vector.

Making an assumption that the non-wage error terms (the ε's) are iid extreme value across all choices and time periods and are also independent of the wage errors reduces the complexity of computing value functions in two ways.[32] First, it removes one source of correlation between C^j_t and S^j_{t+1}. Secondly, it reduces the dimension of the integral in equation (3.9). To see this, note that because the set of wage errors, $\bar{v}^j_{t+1}$, and the set of non-wage errors, $\bar{\varepsilon}^j_{t+1}$, are mutually exclusive and collectively exhaustive with respect to the set S^j_{t+1}, equation (3.9) can be rewritten as:

$$(3.10) \qquad EZ^j(t,X,Y_t,S^j_{t+1}) = \int \int \max^j(\bullet)\, g(\bar{\varepsilon}^j_{t+1})\, d\bar{\varepsilon}^j_{t+1} \quad g(\bar{v}^j_{t+1} | v^j_t)\, d\bar{v}^j_{t+1}.$$

[32]Stern (1997a) suggests that in many cases this distributional assumption may not be very restrictive.

The argument of the density function $g(\bar{v}^j_{t+1}|v^j_t)$ shows that the inter-period relationship between error terms is now due to the correlation between the wage errors in the model. The innermost integral of equation (3.10) represents the expectation conditional on the set of wage errors. Denoting this conditional expectation $EZ^j(t,\bullet,S^j_{t+1}|\bar{v}^j_{t+1})$ and rewriting (3.10) gives:

$$EZ^j(t,\bullet,S^j_{t+1})=\int EZ^j(t,\bullet,S^j_{t+1}|\bar{v}^j_{t+1})\, g(\bar{v}^j_{t+1}|v^j_t)\, d\bar{v}^j_{t+1}. \tag{3.11}$$

The extreme value assumption implies that $EZ^j(t,\bullet,S^j_{t+1}|\bar{v}^j_{t+1})$ j=Eb, N, and H take on computationally convenient closed form solutions (see Rust, 1987 and Berkovec and Stern, 1991).[33] Therefore, the dimensionality of the integral is reduced by d^j_{t+1}, the number of elements in the set $\bar{\varepsilon}^j_{t+1}$.

A further simplification of the correlation structure comes with the assumption that the wage error in non-teaching jobs and new teaching jobs are normally distributed and independent across time:

$$v^N_{t+1}\sim N(0,\sigma^2_N) \tag{3.12}$$

$$v^{Et+1}_{t+1}\sim N(0,\sigma^2_E) \tag{3.13}$$

where σ^2_N and σ^2_E are variances which will be estimated. In this paper, the correlation between C^j_t and S^j_{t+1} comes from the serial correlation of the teaching wage error in a particular teaching job:

---

[33] $$EZ^j(t,\cdot|\bar{v}^j_{t+1})=\tau\{\lambda+\ln\sum_{k\in\bar{D}^j_{t+1}}\exp[\bar{V}^k(t+1,\cdot,C^k_{t+1})/\tau]\} \quad \textit{for } j=Eb,N, \textit{ and } H$$

where λ is Euler's constant, $\tau^2\pi^2/6$ is the variance of the extreme value error which will be estimated, and $\bar{V}^j=V^j-\varepsilon^j$. τ essentially determines the relative importance of wage and non-wage utility in the decision process.

(3.14) $\nu^{Eb}_{t+1}=\rho\nu^{Eb}_{t}+e_{t+1}$ for $t>b$ where $e_t \sim$ iid $N(0,\sigma^2_e)$.

Specifically, the nature of the inter-period correlation is that $\nu^{Eb}_t \in C^{Eb}_t$ is correlated with $\nu^{Eb}_{t+1} \in C^{Eb}_{t+1} \subset S^{Eb}_{t+1}$. Although the AR(1) process for the teaching wage error represents a strong assumption for the teaching wage errors in the model, it does represent a significant relaxation of the traditional assumption of independence. The complexity of the model increases significantly if more general processes, such as those including more lagged values of the teaching wage error on the right hand side of equation (3.14), are used.

4. Numerical Techniques

4.1 Gaussian quadrature approximation

A closed form solution will not exist for equation (3.11) because value functions do not have a closed form with respect to the wage error state variable. Thus, some type of approximation of equation (3.11) must take place. One option is to approximate the multi–dimensional integral by simulation. An alternative, which is used primarily in this paper, is to approximate the integral using Hermite Gaussian quadrature techniques. The Hermite formula is used to approximate an integral which is of the form $\int_{-\infty}^{\infty} f(m)\exp\{-m^2\}$. The idea of the quadrature technique is that a set of p quadrature points, $\{m_i\}_{i=1}^{p}$, and a set of p weights, $\{a_i\}_{i=1}^{p}$, are optimally determined such $\int_{-\infty}^{\infty} f(m)\exp\{-m^2\}\,dm \approx \Sigma_{i=1}^{p} a_i f(m_i)$. The set of quadrature points and weights for the Hermite formulas are presented in tabular form in several books including Stroud and Secrest (1966). The Gaussian quadrature technique may be useful in this computationally intensive context because of its ability to yield high quality approximations with even a small number of points; Gaussian quadrature formulas using p

quadrature points will provide exact solutions of integrals where f(•) is a polynomial of order less than or equal to 2p-1.

The Gaussian quadrature approximation of equation (3.11) can be written as:

$$(4.1) \qquad EZ^j(t,\cdot)=\frac{1}{\pi^{.5(d^j_{t+1}-1)}}\sum_{i=1}^{p^{(d^j_{t+1}-1)}} A_i EZ^j(t,\cdot,S^j_{t+1}|\bar{v}^j_{t+1,i}(v^j_t)) \quad j=Eb,N,and\ H.$$

where d^j_{t+1}-1 is the number of elements in $\bar{v}^j_{t+1}$, $\bar{v}^j_{t+1,i}$ represents the ith realization of the vector $\bar{v}^j_{t+1}$ prescribed by the quadrature formula, and A_i is a product of d^j_{t+1}-1 quadrature weights. Note that the number of elements in this sum is non–linear in the dimension of the integral, d^j_{t+1}-1. In this application, the integral is three dimensional for any of the value functions associated with teaching because it is necessary to integrate over two teaching wage errors and one non–teaching wage error. Therefore, because the majority of estimation time is devoted to computing equation (4.1), increasing the choice of p has a direct and very substantial impact on the amount of computer time that is needed to estimate the model.[34] Thus, the performance of the approximation for different values of p is crucial to the feasibility of the approach and will be explored in section (6.2). Equation A.1 in appendix (A) shows this approximation without the vector notation for the wage errors. For more detail on the Gaussian quadrature technique see Butler and Moffitt (1982), Sickles and Taubman (1986), and Tauchen and Hussey (1991) who have also used this approximation technique in economic applications.

[34]For example, if p=2, the sum contains 2^3=8 elements. If p=5, the sum contains 5^3=125 elements. Evaluations of the value functions in the p=5 case takes roughly fifteen times as long as the p=2 case.

4.2 Solving Value Functions by Backwards Recursion

The approximation in equation (4.1) involves summing over a discrete set of realizations of the wage error vector, $\{\bar{v}^j_{t+1,i}: i=1,2,...\}$. Theoretically, given this discretization, the dynamic programming problem can now be solved by standard backwards recursion for each person at T^*, the last period in which an individual makes decisions. To do this, the deterministic portion of the value functions, $\bar{V}^j=V^j-\varepsilon^j$ j=E, N, and H, must first be solved for all the elements of the *state space* at T^*, which is defined to be the set of all possible combinations of the state variables, Y_{T^*} and $v^E_{T^*}$, that could arise. In this last period of a person's decision–making horizon, the value functions are trivial because there exist no future choices for the individual to make. Once the value functions are solved at T^*, they theoretically can be solved backwards recursively for all elements of the state space for all $t<T^*$ using the recursive descriptions of the value functions in equations (3.6) and (3.7). However, although theoretically possible, this is not computationally practical. The AR(1) specification for the teaching wage error implies that the set of teaching wage errors for which $V^E(\bullet)$ needs to be solved at time t+1 depends in part on the set of teaching wage errors for which $V^E(\bullet)$ is being solved at time t. The specific inter-period relationship between these sets imposed by the Gaussian quadrature approximation implies that, if p is greater than one, the number of values of $V^E(\bullet)$ for which the teaching value functions need to be calculated increases more than exponentially with time.[35] As noted in the introduction and as discussed in more detail in appendix A, from the standpoint of computation this number quickly becomes impractically large.

[35]This number of possible error term realizations is even greater if equation (3.11) is simulated.

To deal with this problem, a smaller set of equally spaced teaching wage errors, or *error points*, is chosen at each period t for each person. The set of all possible combinations of the error points at time t and the possible values of teaching experience at time t form a subset of the state space which will be referred to as the *state grid* at time t.[36] It is then the elements of the state grid, which is potentially much smaller than the state space, for which value functions are calculated by backwards recursion in each period. One complication arises during this process. When computing value functions for the elements of the state grid at time t it will be necessary to approximate value functions at time t+1 associated with points which are not in the t+1 state grid and, therefore, have not been solved by backwards recursion. This occurs because only a very small subset of the possible wage error realizations, v^E_{t+1}, are chosen as error points and used to construct the t+1 state grid.

In order for the non–parametric approach which is adopted in this paper to succeed, it is necessary that all possible wage error realizations for which value functions will need to be approximated at time t+1 lie in the interval between the largest and smallest error points at time t+1. At each time t+1, the set of possible wage error realizations depends on, the parameters of the model, the set of possible wage error realizations at time t (through the relationship between $\bar{v}^j_{t+1}$ and v^j_t in equation 4.1), and the set of simulated wage errors which, as will be discussed in section (5), are needed to compute choice probabilities in period t+1. This implies that, for each guess of the model parameters, a forward recursion process (starting in the first period that a person is observed) which takes into account the above factors must take place to determine the error points for which value functions will be solved by backwards recursion. While this

[36]Technically, the state grid is not a subset of the state space because the equally spaced error points are not chosen directly from the state space.

process does take some thought and programming effort, it does have a direct beneficial consequence. If value functions have been solved for all possible values of Y_{t+1}, the approximation of the value function associated with any possible continuous error term realization can be constructed using a straightforward, non–parametric, one dimensional, linear interpolation which involves two surrounding points.[37] A discussion of this interpolation and a more in depth discussion of state space issues is given in appendix (A). A desirable property of this solution technique, which was discussed in the introduction, is that the interpolation error can be made arbitrarily small by an appropriate reduction of the spacing, Δ, between error points. Appendix (B) establishes the desirable convergence properties of the Gaussian quadrature approximation in the presence of this interpolation process.[38]

Decreasing the spacing between error points, Δ, increases the amount of computer time necessary to estimate the model because it increases the number of points in the state grid. From a feasibility standpoint, what is important is how the approximation performs for reasonably large values of Δ. This is explored in section (6.2).

5. Econometric Methods

The value functions discussed in the previous section are used in calculating the choice probabilities which enter the simulated maximum likelihood routine used to estimate the parameters of the model. Recall that there are several sources of information which are known

---

[37]To interpolate the value function $V^E(\bullet, Y_{t+1}^*, \nu_{t+1}^{E*})$, a weighted average of $V^E(\bullet, Y_{t+1}^*, \nu_{t+1}^{E\prime})$ and $V^E(\bullet, Y_{t+1}^*, \nu_{t+1}^{E\prime\prime})$ is formed where $\nu_{t+1}^{E\prime}$ is the smallest error point which is larger than ν_{t+1}^{E*} and $\nu_{t+1}^{E\prime\prime}$ is the largest error point which is smaller than ν_{t+1}^{E*}.

[38]The proof is very similar if the expected future utility integral is approximated by simulation rather than Gaussian quadrature methods.

to the individual when making decisions but may not be observed by the researcher. First, the wage offers from options which are not chosen by the individual are not observed in the data, and, as mentioned in the data section, the survey does not provide wages for the accepted job in every year that the person works.[39] Second, the individual knows his heterogeneity components, μ^E,μ^N,μ^H, which are unobserved by the econometrician. Finally, the individual is assumed to observe the current period non–pecuniary error terms associated with each of the choices available to him.

Notationally, let $\bar{v}_t^{j,m}$ represent the set of all relevant wage errors which are unobserved by the econometrician at time t (conditional on the choice of j in the previous period) and let $\mu=\{\mu^E,\mu^N,\mu^H\}$.[40] The assumption that the non–pecuniary error terms have an extreme value distribution implies that conditional on knowing the values of $\bar{v}_t^{j,m}$ and μ, the probability of person i making choice D in time t takes on a convenient closed form solution:

$$(5.1) \qquad P_{i,t}(D,X,Y_t,S_t^j|\bar{v}_t^{j,m},\mu) = \frac{\exp[\bar{V}^D(t,X,Y_t,C_t^D)/\tau]}{\sum_{k\in\bar{D}_t^j}\exp[\bar{V}^k(t,X,Y_t,C_t^k)/\tau]} \quad \textit{for } j=Eb,N,H$$

where as before, $\bar{V}^k=V^k-\varepsilon^k$. (Note that the potential elements of $\bar{v}_t^{j,m}$ are the current period wage errors associated with each choice k, v_t^k, which enter value functions as elements of C_t^k. Note also that μ^k is an implicit argument of $\bar{V}^k$).

---

[39]Recall that the starting wage in a new teaching job is always observed. Typically a subset of the subsequent wages are also observed.

[40]Thus, $\bar{v}_t^o=\{v_t^D\}$ if a wage is observed and $\bar{v}_t^o$ is the empty set if a wage is not observed in the accepted job..

If a wage is not observed at time t, the likelihood contribution for person i in period t is given by the choice probability in equation (5.1). However, if a wage error, v_t^D, is observed, the likelihood contribution is the joint probability of the observed wage and the choice that the person makes. This can be written as the product of the probability of the observed wage and the choice probability conditional on the observed wage. Thus, again conditioning on the information, $\bar{v}_t^{j,m}$ and μ, which is not observed by the econometrician, the likelihood contribution for person i at time t is given by:

$$L(i,t|\bar{v}_t^{j,m},\mu)=P(v_t^D|v^D_{t_o})\ P_{i,t}(D,X,Y_t,S_t^j|\bar{v}_t^{j,m},\mu) \quad \text{for } j=Eb,N,H \tag{5.2}$$

where t_o is the most recent previous year in which a wage was observed for person i. The serial correlation allowed in teaching wages necessitates conditioning the current period wage probability on the last observed wage. Note that the most recently observed wage will not affect the distribution of v_t^D if time t is the first year at a particular teaching job or the current job is in a non–teaching occupation.[41]

The likelihood contribution for person i conditional on all missing wages and μ is the product of the period specific likelihood contributions over the years in which a person is observed.[42]

---

[41]Non–teaching wages and wages in different teaching jobs are assumed to be serially uncorrelated. The probability for starting teaching jobs is computed using the wage error distribution in equation (3.13). The advantage of always having starting wages in a teaching job is that a previous wage error always exists to condition on when computing wage probabilities.

[42]For reasons of notational simplicity, the j superscript on $\bar{v}_t^{j,m}$ is no longer written explicitly.

(5.3) $$L(i|\bar{v}^m,\mu)=\prod_{t=1}^{\bar{T}} L(i,t|\bar{v}_t^m,\mu)$$

where $\bar{v}^m$ is $\bar{v}_1^m \cup \bar{v}_2^m \ldots \cup \bar{v}_T^m$. The unconditional likelihood contribution for person i involves integrating equation (5.3) over the joint distribution of the unobserved heterogeneity and the joint distribution of the unobserved wages conditional on the set of observed wages, $\bar{v}^o$.

(5.4) $$L(i)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L(i|\bar{v}^m,\mu)\ \ g(\bar{v}^m|\bar{v}^o)\ h(\mu)\ d\bar{v}^m\ d\mu.$$

This integral can be simulated as:

(5.5) $$L(i)\approx\sum_{k=1}^{K}\sum_{r=1}^{R} L(i|\bar{v}^m{}_k,\mu_r)$$

where $\bar{v}^m{}_k$ represents the k'th of K draws of $\bar{v}^m$ conditional on $\bar{v}^o$, and μ_r represents the r'th of R draws of μ.[43]

The simulations of μ take into account the following assumption about the distribution of unobserved heterogeneity:

(5.6) $$\mu^E=\sigma_1\psi_E,\ \mu^N=\sigma_2\psi_N,\ \text{and}\ \mu^H=\sigma_3\psi_H.$$

ψ_E, ψ_N, and ψ_H are independent, identically distributed standard normal random variables, and σ_1, σ_2, and σ_3, are the standard deviations of the unobserved heterogeneity distribution which measure the relative importance of unobserved differences between people in the model.

[43]Another possibility would be to simulate the integral using a single summation. The technique in equation (5.5) is more sensible because estimation time increases almost linearly in number of draws of unobserved heterogeneity, but only slightly in the number of wage draws. The reason for this is that a change in heterogeneity values requires resolving value functions whereas this is not the case for changes in wage draws.

$\bar{v}^m$ will consist of both teaching and non–teaching wage errors. Simulating unobserved values of the non–teaching wage error is straightforward given the independence of non–teaching wages in the model. The need to condition $\bar{v}^m$ on $\bar{v}^o$ comes from the serial correlation of teaching wages within a particular teaching job. Given the AR(1) wage process, the distribution of the unobserved wages in any teaching job conditional on the observed wages in that job have a joint normal distribution with an easily computable mean and covariance matrix. The missing wages can be simulated in a straightforward fashion from this distribution. See Stern (1997b) for an overview of simulation methods.

The goal is to determine the parameter values σ_E, σ_N, σ_e, ρ, τ, σ_1, σ_2, σ_3, and the parameters in $\bar{W}^i(\bullet)$ and $\bar{Q}^i(\bullet)$ i=E and N which maximize the log likelihood function in equation (5.8). Coefficients are not estimated for the home option (H) because it was selected as the base case.[44] The analytical derivatives of the likelihood function with respect to each of the model parameters can be computed. The basic building blocks of these analytical derivatives are the derivatives of the person–specific value functions with respect to the model parameters. In general, the derivatives of the value functions are continuous in this application. However, this is not true at the error points themselves, where, although the value functions are continuous, the derivatives from the left are not equal to the derivatives from the right.[45] One option in this case

[44]Discrete choice models can only identify preferences relative to a base case. Therefore, the coefficients associated with the home option (H) were fixed at zero.

[45]Suppose that v^{E*}, v^{E**}, v^{E***} are three consecutive error points. Let γ be small. Then, $v^{E**}+\gamma$ is interpolated using v^{E**} and v^{E***} but $v^{E**}-\gamma$ is interpolated using v^{E*} and v^{E**}. The derivative of $V(\bullet,v^{E**}-\gamma)$ with respect to any parameter is essentially a simple weighted average of the derivatives of $V(\bullet,v^{E*})$ and $V(\bullet, v^{E**})$ with respect to that parameter, whereas the derivative of $V(\bullet, v^{E**}+\gamma)$ is a simple weighted average of the derivatives of $V(\bullet, v^{E**})$ and $V(\bullet, v^{E***})$. A discontinuity in the derivatives arises at v^{E**} because the derivative of $V(\bullet,v^{E*})$ continues to influence the derivative of $V(\bullet,v^{E**}-\gamma)$ even as γ approaches zero, and the derivative of $V(\bullet,v^{E***})$ continues to influence the derivative of $V(\bullet,v^{E**}+\gamma)$ even as γ approaches zero.

is to simply abandon the use of derivative based updating algorithms altogether. However, given the substantial computational burden of these types of models and the significant computational cost difference between derivative and non–derivative based approaches, it makes more sense to start with a derivative based method and to switch to a non–derivative based method when the updating algorithm "eventually" reaches an iteration at which changing the parameters in the direction that is prescribed by the updating algorithm does not lead to an improvement in the likelihood function.[46] When this approach was taken here, it was found that the Newton–Raphson updating algorithm (with analytical derivatives) typically did not get "stuck" until a reasonable convergence criteria had been satisfied. Further, robustness checks showed that the estimates that are achieved are robust to the starting values which are used.[47] Thus, it seems that the derivatives of the likelihood function are quite well–behaved despite the fact that the derivatives of the person–specific value functions have potential discontinuities. If the spacing between error points, Δ, is chosen to be large, this is likely to be the case because relatively few

---

[46] Derivative based methods are much faster than non–derivative based methods. Further, using analytical derivatives rather than numerical derivatives reduces estimation time in this application by a factor of six or seven.

[47] For example, for the Δ=.1 case, the model was estimated using four very different sets of starting points. At the iteration where the likelihood function failed to improve, the square of the derivative of the log likelihood function with respect to each of the parameters was computed. Given standard assumptions about the concavity of the likelihood function, the conditions for convergence are satisfied when these derivatives are equal to zero.

For the parameter values associated with three of the four sets of starting points, the average of the squared derivatives was less than .000002 at the iteration where the likelihood function failed to improve.. Further, the likelihood values and point estimates were virtually identical between these three sets. For example, for the two sets of parameters that were furthest apart, the average parameter estimate from one set differed from its counterpart in the other set by only .0004 of its estimated standard error.

For the fourth set of estimates, the average of the squared derivatives of the log likelihood function with respect to each of the parameters was approximately .27 at the iteration where the log likelihood function failed to improve (the average derivative of the log likelihood function was slightly less than .1). However, even in this case, the estimated parameters were very similar to the estimated parameters from the cases above. Specifically, the average parameter differed from its counterparts in the other three cases by only about .005 of its estimated standard error.

of the v^E values for which value functions are calculated during model solution or estimation are "close" enough to the error points to be directly affected by the discontinuities. If the spacing between error points, Δ, is chosen to be small, this is likely to be the case because the difference in the derivatives of the value functions from the left and the derivatives of the value functions from the right at the error point will be small. In other applications, the Newton–Raphson updating algorithm might not perform as well as it does in this application. However, Brien, Lillard, and Stern (1996) show that a very minor modification of the interpolating function used here will remove the discontinuities discussed above while maintaining the spirit and properties of the local interpolating approach which is proposed, implemented, and tested in this paper.[48]

6. Results

6.1 Specification of Deterministic Portions

The deterministic portions of the utility equations, $\bar{W}^j(\bullet)$ and $\bar{Q}^j(\bullet)$ for j=E and N, are assumed to be linear functions of their arguments. The variables that appear in either $\bar{W}^j(\bullet)$ or $\bar{Q}^j(\bullet)$ are: a constant, CONST, the time trend (1=1975,..., 1986=11), TIME, the time trend squared, TIMESQ, a dummy variable indicating whether the person is male, MALE, the math SAT score divided by 100, SAT, the years of post-bachelor education accumulated by time t, EDU, the teaching experience accumulated by time t, EXP, the number of children at time t,

[48]Suppose we are interpolating $v^{E*}+\gamma$ using the surrounding error points v^{E*} and v^{E**}. Then, under their interpolating function the interpolating weight associated with v^{E**} would be $\gamma^\varphi/[\gamma^\varphi+(v^{E**}-v^{E*}-\gamma)^\varphi]$ where φ is some positive constant. The weight associated with v^{E**} would be defined similarly. In general, for $\varphi \geq 0$ (excluding $\varphi=1$) the derivatives of the likelihood function are continuous at the error points. Essentially, this occurs because as γ becomes small, the derivative of $V(\bullet,v^{E*}+\gamma)$ with respect to any parameter that influences γ depends only on the derivative of $V(\bullet,v^{E*})$ with respect to that parameter. The simple weighted average used in this paper is a special case of their interpolating function in which $\varphi=1$.

CHILD, and a dummy variable indicating whether a person is married at time t, MARR. Although it would be desirable to estimate the model separately for males and females, this is not done because of the relatively small sample. However, in order to take into account that some variables are very likely to have different effects for male and females, the interaction variables MALExCHILD and MALExMARR are included.[49] It is important to note that, to the extent that individuals make marital decisions, fertility decisions, and work decisions jointly, it is not truly correct to model the marital and fertility variables exogenously. Van der Klaauw (1996b) estimates a dynamic programming model of the joint marital and work decisions.

6.2 Examination of approximation quality.

The amount of time necessary to estimate the model with unobserved heterogeneity increases linearly in the number of simulation draws which are used to approximate the unobserved heterogeneity integrals in equation (5.4). Therefore, the success of the general approach discussed above depends critically on the performance of the approximation when the number of quadrature points, p, and the spacing between error points for which value functions are actually solved by backwards recursion, Δ, are set by the econometrician at reasonable levels. By using the model described above, but without heterogeneity, it is computationally feasible to examine the robustness of estimates to both components of the approximation.[50] Taking into account the tradeoff between precision and computational cost, these findings can then be used

[49] The effect of a person's race was not estimated due to the small number of minorities in the data.

[50] For the purposes of the approximation, the model is also estimated without the interaction terms involving MARR and CHILD because these terms substantially increase the number of iterations that are needed in order to reach convergence.

to choose reasonable values of p and Δ for the estimation of the model with unobserved heterogeneity.

In order to understand the implications of the following results on the general feasibility of the approach, what really matters is the size of Δ relative to the standard deviation of the serially correlated unobservable. In this application, the estimated standard deviation of starting wages, σ_E, is approximately .35. While the exact solution can never be obtained for any finite choice of p or Δ, the results will show that it is reasonable to believe that the parameter estimates from the p=6 and Δ=.1 case are very close to the "true" estimates. Therefore, it seems reasonable to examine approximation quality by comparing the difference between these parameter estimates and estimates which are obtained for smaller values of p or larger values of Δ. For expositional purposes, the difference between a particular parameter and its p=6, Δ=.1 counterpart will be referred to as the approximation bias of the parameter. For a particular parameter, it will generally be more informative to examine the approximation bias as a proportion of its estimated standard error than the magnitude of the bias itself.[51] The average (of the absolute value) of this weighted approximation bias over all the parameters in the model is computed and used to facilitate comparisons of approximation quality over entire sets of parameters.

Holding Δ constant at a value of .1, figure (3) shows the statistic for different levels of p. Notice that the average estimated parameter changes by only .01 of its standard error if p is decreased from p=5 to p=6. Thus, it is reasonable to conclude that, ignoring the effect of changing Δ, the parameters obtained using p=6 are very close to the true values. What is

[51]The researcher is likely to be willing to accept approximation bias which is small relative to the size of the variation due to sampling.

important to note about figure (3) is that the approximation works very well for small values of p. At p=3, the average estimated parameter differs from its p=6 counterpart by less than .03 of its standard error. The approximation remains quite good even at p=2 where the approximation bias statistic is .11. These results are very important from the standpoint of feasibility given the direct relationship between p and the amount of time which is necessary to estimate the model.[52]

Given the tradeoff between time savings and approximation quality, the choice of p=3 seems logical for the estimation of the model with unobserved heterogeneity. The second dimension that must be chosen by the econometrician is the value of Δ. Figure (4) shows the value of the approximation bias statistic for many choices of Δ where, in order to isolate the effect of changes in Δ, the approximation bias is measured relative to the parameter estimates which are obtained using p=3 and Δ=.05. At Δ=.05, points are spaced only about .14 of a wage standard deviation apart. When Δ is increased from .05 to .1, the average parameter changes by only .01 of its standard error. Therefore, it seems reasonable to assume that the choice of Δ=.1 gives estimates which are very close to the true estimates for a given value of p. What is important is that the approximation remains accurate for large values of Δ. For example, if Δ is quadrupled from .05 to .2, the approximation bias for the average estimate is only .04 of its standard deviation. At Δ=.4, the error points are spaced more than one standard deviation apart, but the approximation bias statistic remains less than .1.

[52]Recall that the number of elements in the quadrature sum used to approximate the expected future utility component of any teaching value functions and any non–teaching value functions are p^3 and p^2 respectively. This almost directly determines estimation time. For example, in the model (without heterogeneity) with p=3 and Δ=.1 it takes approximately 18.6 minutes to evaluate the likelihood function and analytical derivatives associated with each update of the Newton–Raphson, whereas in the model with p=6 and Δ=.1 it takes approximately 2.33 hours per update.

Further, results indicate that approximation quality can be improved at a given computational cost by not spacing error points evenly. When error points are spaced an average of .4 apart, but the spacing is closer for points which are more likely to have a direct effect on the likelihood function, the approximation bias statistic is less than .06 (compared to .10 for the Δ=.4 case).[53] When error points are variably spaced an average of .8 (2.29 wage standard deviations) apart, the approximation bias statistic is .26 (compared to .39 for the evenly spaced Δ=.8 case). Intuitively, the variable spacing method will be most important in applications in which a large portion of the possible error realizations have a low probability of actually influencing the likelihood function. This occurs in this application because the estimated value of ρ is approximately .91 which implies that the range of possible error realizations becomes larger over time.[54] The first column of table (2) shows the average (over all people) range of possible error realizations in each period given the p=3, Δ=.1 estimates. With ρ=.91, the range of possible error realizations is greater than 28 wage standard deviations in some periods. Clearly, many of these "possible" realizations will not have much of an impact on the actual likelihood function.

Recall that the ultimate goal is to estimate the model with heterogeneity. The approximation quality results suggest that there are many different possible combinations of p and Δ that could feasibly provide estimates of the heterogeneous model with only small amounts

---

[53]Essentially this means that error points are spaced further apart as they become closer to the minimum and maximum of the range of possible values.

[54] Appendix A shows that the set of time t+1 wage realizations that must be available to solve the time t value function associated with the largest time t error point v^* is $\{\sqrt{2}\sigma_\varepsilon m_z+\rho v^*: z=1,2,...,p\}$. A high value of ρ will imply that the largest value of this set is greater than v^*, in which case the range of possible errors at time t+1 will be larger than the range of possible errors at time t. If ρ is relatively small, the largest element of the set will be smaller than v^* and the range of possible errors may become smaller over time.

of approximation error. It was decided to use three quadrature points (p=3) and the variable spacing procedure with an average spacing equal to .4 (avgΔ=.4). To get a more concrete sense of approximation quality for the p=3, avgΔ=.4 case, table (3) shows these estimates, the p=6, Δ=.1 estimates, the approximation bias, the estimated standard error of the p=6, Δ=.1 estimates, and the approximation bias for each estimate as a proportion of its standard error. What is important is that, although the p=3, avgΔ=.4 choice requires only .033 times as much computational time as the p=6, Δ=.1 case, the average approximation bias relative to the p=6, Δ=.1 model is only .05 of its standard error.

6.3 Estimates of Full Model

Table (4) shows the coefficient estimates from the heterogeneous model using p=3, avgΔ=.4, and assuming a discount factor, β, of .95. The coefficients in the wage equations measure the average effect that increasing a particular variable has on real weekly log wages. The numbers in parentheses represent the asymptotic standard errors. Not surprisingly given the rigidity of the traditional wage structure, the post-bachelor education level and the years of teaching experience are significant determinants of average teaching wages. The negative coefficient on the year and the positive coefficient on the year squared variable represent a significant u shaped time trend for teaching wages. This is consistent with previous literature which has shown that real teaching wages decreased substantially after 1972 until beginning a recovery in the early 1980's.[55]

[55]See the Condition of Education, 1994. The National Center for Education Statistics, U.S. Department of Education. Since everyone in the sample graduated from high school in the same year, the effect of age cannot be differentiated from the time trend.

Table (4) indicates that, on average, individuals with high math SAT scores do not receive higher wages in teaching jobs but do receive significant wage premiums in non–teaching jobs. Thus, these estimates support the notion that lower teaching participation rates for academically gifted teachers stem in part from the ability of these teachers to obtain higher than average wages in the non–teaching sector but not in the teaching occupation. The model estimates indicate that the non–teaching wage premium does not explain all of the differences in the teaching participation rate across SAT groups. For example, the coefficient on SAT in the non–pecuniary teaching equation, –.083, shows that individuals with higher SAT scores enjoy teaching significantly less than other teachers (relative to the option of not working).[56]

Table (4) shows that previously unavailable marital and fertility variables play a very important role in determining individual decisions. The estimate of CHILD, which represents the effect of children for females, indicates that women with children are significantly less likely to be working in either teaching jobs or non–teaching jobs than women without children. However, the effect of children for males (CHILD+CHILDXMALE) is small and insignificant in both the teaching and non–teaching non–pecuniary equations. Similarly, being married makes a woman much less likely to work but has a smaller effect on males.[57]

6.4 Policy Simulations

The effects of two changes in the teacher wage structure are now examined. The first, denoted "Policy one", corresponds to the often heard normative statement that "teachers should

[56]Stinebrickner (1996a) suggests that this effect may in part be due to the omission of non–teaching experience.

[57]If individuals who are unhappy with their jobs a in unobservable ways are more likely to leave work in order to have children, the effect of children on job exits would be overstated. In this scenario, the people with children would be more likely to leave teaching even if they did not have children. .

earn more money." It involves a uniform wage increase, of 25% on average, for all teachers under the traditional, rigid wage structure. The second policy, "Policy two", explores the possibility of implementing a wage policy which deviates from the traditional wage structure. It raises the average teacher wage by 25% but does so by basing the amount of the raise on the ability level of the teacher. In particular, the wage increase that an individual receives depends linearly on the person's SAT score with the lowest SAT person receiving no additional wage.[58]

The details of the simulation procedure are discussed in Appendix C. The procedure is first performed on the unchanged set of estimates from the model to get a baseline set of results, and then performed on two sets of parameters which have been modified to be consistent with policy one and policy two. Figure (5) shows that Policy one and Policy two have virtually identical effects on the labor supply of the overall sample. The increase in the proportion of aggregate years that individuals choose to teach is due to both an increase in the number of people who actually enter teaching at some point, and, as shown in figure (6), an increase in the length of spells for those who choose to enter teaching.

An important goal of this work is to examine how different policies affect different types of teachers.[59] Figure (7) shows a measure of the labor supply for people in the top 1/3 of the

---

[58]The original log wage equation can be written as W_1=CONST + α(SAT)+... For Policy two, the goal is to find a new wage equation which implies and average wage (not log) increase of 25% with the lowest ability person receiving no wage increase. The minimum math SAT score observed in the data is 2.70. Then, the goal is to find the value of A in the equation W_2=CONST + α(SAT)+ A(SAT-2.70) +... such that the average value of $[\exp(W_1)-\exp(W_2)]/\exp(W_1)$ is .25.

[59]Ballou and Podgursky (1996) find that raising wages is not a sufficient condition for improving the teaching profession even if higher wages are successful at attracting better new applicants. They argue that since labor markets are often in a state of excess supply, the effectiveness of wage increases depends on the composition of the (otherwise exiting) teachers who are induced to stay when wages are increased. This paper does not examine the decision to become certified to teach, but does allow an examination of the effects of wage increases on current teachers of varying ability.

sample in terms of SAT scores *relative* to that of individuals in the lowest 2/3 of the sample in terms of SAT scores. In particular, it shows the aggregate proportion of periods that each option is chosen by the low ability group divided by the aggregate proportion of periods that the option is chosen by the high ability group. For example, the ratio for the teaching option under the unchanged baseline wage structure, .88, is found by dividing the aggregate teaching proportion of the high group, .42, by the teaching proportion of the low group, .48. Comparing this number to the ratio from policy one, .91, (proportion/high=.50 proportion/low=.55 percent) shows that a uniform wage leads to only a very small increase in the proportion of time that the high ability group spends teaching relative to that of the low group. However, notice that this improvement is greater for policy two. The implementation of policy two leads to a ratio for the teaching option of 1.01 (high=.53,low=.53). Thus, the simulations indicate that policy two induces the high ability teachers to choose teaching as often as the low ability teachers.[60] The difference in the ratios between Policy one and Policy two in figure (7) is due in part to the relative effect of the policies on participation and due in part to the relative effect of the policies on duration.[61]

---

[60]This paper examines the participation decision of certified teachers but does not examine the decision to become certified. Ballou and Podgursky (1995) find that even if raising wages leads to more qualified job applicants, raising wages is not a sufficient condition for improving the teaching profession. They argue that since labor markets are often in a state of excess supply, the effectiveness of wage increases depends on the composition of the (otherwise exiting) teachers who are induced to stay when wages are increased. This paper suggests that the type of wage increase matters from this standpoint.

[61]Policy one leads to only a very slight decrease in the difference between the survivor functions for the high ability and low ability individuals. The survivor function of the low ability group is still always greater than the survivor function of the high ability group. Under Policy two, the survivor function of the high ability group is above the survivor function of the low ability group. The participation rates of the high group in the baseline simulation, policy one, and policy two are .73, .79, and .82 respectively. For the low group the participation rates are .76, .82, and .81 respectively. Thus, policy two leads to an increase in the participation rate of the high group relative to that of the low group when compared to policy one and the baseline policy.

7. Conclusion

This paper presents a new approach for allowing serial correlation in dynamic, discrete choice models. The approach centers around a new, non–parametric value function interpolation algorithm. The nature of this solution method ensures that the econometrician has complete control over the factors which ensure that parameter estimates from an approximate solution to a model with serial correlation can be made arbitrarily close to the true parameter estimates.

Contrary to the prevailing belief, the results of this paper suggest that the estimation of dynamic, discrete choice models with serial correlation can be achieved with little approximation bias, even without incurring large amounts of computational costs. There are two important findings in this application that lead to this conclusion. First, the results suggest that Gaussian quadrature methods with small numbers of quadrature points can yield accurate approximations of the integrals which are necessary to account for the impact that future uncertainty about the realizations of serially correlated unobservables has on the future utility that a person expects. For example, in this application, approximation bias remained reasonably low even when three dimensional integrals are approximated using quadrature sums with as few as eight total elements. In many other applications which include a single, serially correlated error term, only a single dimensional integral would have to be evaluated.[62] The results suggest that this could be accomplished reasonably well using a quadrature sum with as few as only two or three elements.

Secondly, the number of possible error values for which the value functions need to be actually solved by backwards recursion is potentially quite low. In this application, the

[62]For example, this would be the case in a retirement model if we wanted to model health status as a continuous, serially correlated variable. What differs in the application in this paper is that a person gets a wage offer from up to three potential employers.

approximation remained quite accurate when these points were spaced in excess of one standard deviation apart and the value functions associated with all other elements of the state space were interpolated using the non–parametric approach. This application was still computationally burdensome because, as table (2) indicated, the range of possible error term realizations was quite large. This was directly related to the large value which was found for the serial correlation coefficient in this particular application. The second and third columns of table (2) show the possible range of error terms that would have to be solved if the serial correlation coefficient was .7 or .5 instead of .91. In the hypothetical ρ=.5 case, the range of possible error realizations is constant at about 2.2. Thus, the results suggest that the approximation would remain quite accurate in this hypothetical case if value functions were solved for only five or six possible values of the wage error state variable. It is important to note that these results also have more general implications on the ability of researchers to closely approximate the solution to dynamic programming models when continuous variables are present, regardless of whether these variables are serially correlated.

The general approach taken in this paper can be generalized to more than one serially correlated error term. Appendix D briefly examines one easily implementable interpolating function for multiple dimensions which is based on a local ordinary least squares regression approach. However, what appears to be useful is the general approach which limits the influence of functional form assumptions by basing interpolations on points which "surround" the point for which value functions are being solved. The interpolating function recently proposed by Brien, Lillard, and Stern (1996) for two continuous variables (and applicable in greater dimensions) is, in some sense, more consistent with the approach taken in this paper when one

serially correlated unobservable is present and has properties which make it more desirable than the ordinary least squares regression approach from the standpoint of estimation.

In general, it is likely to be the case that the quality of any approximation method will vary based on a researcher's particular application. Hopefully, the reduction in reliance on functional form assumptions will increase the robustness of approximation quality across applications. Nonetheless, one should naturally be very cautious when trying to conclude how the results found here will generalize to other applications. However, one of the desirable aspects of this approach is that the researcher can often gauge the success of the method in his own application.

Appendix A: State Space Issues

The goal of this section is to more thoroughly examine the state space issues involved with solving the value functions in the model. To facilitate this, the arguments of the value functions which are not state variables will be suppressed.

At each time t, the value function $\bar{V}^E(t,Y_t,v^{Eb}_t)$ must be solved for all combinations of Y_t and all possible teaching wage error values, v^{Eb}_t, that could arise from either old jobs (b<t) or new jobs (b=t). As of period t, the possible years of teaching experience that the individual could have acquired is given by the set {0,1,2,...,t-1}. The relevant number of possible values of v^{Eb}_t is not as straightforward to calculate. To see this, consider using equation (3.6) to calculate the value of teaching, $\bar{V}^E(t,Y_t,v^{Eb}_t)$, in a job with wage error v^{Eb}_t. Removing the vector notation, the Gaussian quadrature approximation of the expected future utility component of (3.6) shown originally in equation (4.1) can be written as:

$$(A.1)\qquad EZ^E(t,\cdot)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} EZ^E(t,\bullet\,|v^{Eb}_{t+1},v^{Et+1}_{t+1},v^{N}_{t+1})\; g(v^{Eb}_{t+1},v^{Et+1}_{t+1},v^{N}_{t+1})\; dv^{Eb}_{t+1}\; dv^{Et+1}_{t+1}\; dv^{N}_{t+1}$$

$$\approx \frac{1}{\pi^{1.5}}\sum_{k=1}^{p} a_k \sum_{y=1}^{p} a_y \sum_{z=1}^{p} a_z EZ^E(t,\bullet\,|\sqrt{2}\,\sigma_e m_z+\rho v^{Eb}_t,\sqrt{2}\,\sigma_E m_y,\sqrt{2}\,\sigma_N m_k).$$

Thus, approximating $EZ^E(t,\bullet)$ requires the knowledge of $\bar{V}^E(t+1,Y_{t+1},v^{Eb}_{t+1})$ for the p values of v^{Eb}_{t+1} in the set $\{\sqrt{2}\sigma_e m_z+\rho v^{Eb}_t: z=1,2,...,p\}$. A repetition of this argument reveals that the calculation of $\bar{V}^E(t+1,Y_{t+1},v^{Eb}_{t+1})$ for each of these p values of v^{Eb}_{t+1} requires the knowledge of $\bar{V}^E(t+2,Y_{t+2},v^{Eb}_{t+2})$ for p values of v^{Eb}_{t+2}. In general, the number of values of v^{Eb}_{t+i}, which must

be calculated in order to calculate the value function $\bar{V}^E(t,Y_t,v^{Eb}_t)$, grows exponentially with i and is given by $(p)^i$.[63]

Due to the exponential growth in the state space caused by the serial correlation, the amount of computer time necessary to solve value functions, for all elements of the state space, by the standard backwards method becomes infeasible. The interpolation algorithm which is implemented to deal with this problem involves, for each time t, specifying a *state grid*, which is the combination of the possible values of Y_t with a smaller set of wage errors, or *error points*, at time t.[64] To construct the grid at some time t, the discussion in the previous paragraph is taken into account and the smallest error value, v^L_t, and the largest error value, v^U_t, that could arise in period t are calculated.[65] It then follows that for any possible value of v^{Eb}_t that could arise in period t, $v^{Eb}_t \varepsilon [v^L_t, v^U_t]$. This interval is then discretized into the finite number of equally spaced error points.[66] More specifically, let R^v_t represent the set of error points at time t, and let Δ be the spacing between error points which is set by the econometrician. Then, the set of error points at time t can be written as:

$$R^v_t = \{[v^U_t - v^L_t]/2 \pm j*\Delta: \ j=1,2,...,B\} \tag{A.2}$$

---

[63]Thus, the set of possible teaching wage errors at time t depends in part on the set of possible teaching wage errors in time t-1. It also depends on certain estimation issues such as the wage errors that are observed in accepted jobs at time t and the appropriate distributions of wage errors needed to adjust for all rejected wage offers and any unobserved accepted wage offers.

[64]The set error points at time t, is chosen in a way which ensures that 1) the number of points is small enough that it is feasible to solve value functions by backwards recursion for all combinations of the elements of the state grid and 2) it is possible to interpolate any necessary value functions associated with combinations of state variables which are not in the state grid.

[65]Since the possible error values at time t depend in part on the possible error values in time t-1, v^L_t and v^U_t depend in part on v^L_{t-1} and v^U_{t-1}. This necessitates starting at the first period and working forward when constructing the state grid.

[66]Engen (1991), Palumbo (1991), and Christensen (1990) all deal with discretizing continuous variables.

where B is the smallest integer for which $[v^U_t - v^L_t]/2 + B*\Delta > v^U_t$. This ensures that the largest element of R^v_t is greater than v^U_t and that the smallest element of R^v_t is smaller than v^L_t. Let R^Y_t denote the set of all possible values of teaching experience that the individual could have accumulated as of time t. Then R^Y_t is given by:

(A.3) $$R^Y_t = \{0,1,2,...,t-1\}.$$

Let r^v_t and r^Y_t be scalars which represent elements of R^v_t and R^Y_t respectively. The state space grid, R_t, at time t, is the Cartesian Product of R^v_t and R^Y_t:

(A.4) $$R_t = R^v_t \times R^Y_t = \{(r^v_t, r^Y_t) \mid r^v_t \varepsilon R^v_t,\ r^Y_t \varepsilon R^Y_t\}.$$

Once the state grid has been determined for all time periods, value functions can be solved for all elements of the state grid (for all time periods) by backwards recursion. However, to do this will require some interpolation. Suppose that the value of teaching, $\bar{V}^E(t, r^Y_t, r^v_t)$ is being solved for some grid point (r^v_t, r^Y_t). Let v^{Et}_{t+1} represent a wage error at time t+1 associated with returning to the job characterized at time t by the error point r^v_t. Then as discussed earlier, the calculation of $\bar{V}^E(t, r^Y_t, r^v_t)$ requires the knowledge of $\bar{V}^E(t+1, r^Y_t+1, v^{Et}_{t+1})$ for the p values of v^{Et}_{t+1} given by $\{\sqrt{2}\sigma_e m_z + r^v_t : z=1...p\}$.[67] However, since value functions have not been calculated at time t+1 for all values of the wage error that could arise but, rather, only a certain set of error points, $\bar{V}^E(t+1, r^Y_{t+1}, v^{Et}_{t+1})$ will not have been calculated for any of the p values in this set.[68] Let $\hat{v}^{Et}_{t+1}$ represent any one of the p values of v^{Et}_{t+1}. Then $\bar{V}^E(t+1, r^Y_t+1, \hat{v}^{Et}_{t+1})$ is interpolated as a weighted average of $\bar{V}^E(t+1, r^Y_t+1, r^{v,1}_{t+1})$ and $\bar{V}^E(t+1, r^Y_t+1, r^{v,2}_{t+1})$ where $r^{v,1}_{t+1}$ and $r^{v,2}_{t+1}$ are the closest two error points to $\hat{v}^{Et}_{t+1}$ such that $r^{v,1}_{t+1} < \hat{v}^{Et}_{t+1} < r^{v,2}_{t+1}$. Notice that through a reduction of

---

[67] It also requires $\bar{V}^E(t+1, r^Y_t+1, v^{Et+1}_{t+1})$ for the p values of v^{Et+1}_{t+1} given by $\{\sqrt{2}\sigma_E m_y : y=1..p\}$.

[68] Getting an error value that exactly coincides with one of the error points is a zero probability event.

the spacing, Δ, between error points, $r^{v,1}_{t+1}$ and $r^{v,2}_{t+1}$ can be made arbitrarily close to $\hat{v}^{Et}_{t+1}$. Thus, this source of interpolation error can be made arbitrarily small. The price of this interpolation error reduction is the additional computational burden which is incurred due to an increase in the number of points in the error set.

Appendix B: Convergence of the Gaussian Quadrature Approximation

In this section, the convergence properties of the quadrature approximation of $EZ^E(t,\bullet)$, $EZ^N(t,\bullet)$, and $EZ^H(\bullet)$ are examined. For illustration, consider equation (A.1) which shows $EZ^E(t,\bullet)$ and its Gaussian quadrature approximation. It is desirable to show that by controlling the number of quadrature points, p, and the size of the increment between error points, Δ, the econometrician can ensure that the approximation of $EZ^E(t,\bullet)$ is arbitrarily close to the true value. This will be shown here for the innermost integral of equation (A.1). Realizing that the elements of g are independent, normal random variables, the innermost equation can be written in the form:

$$Z^{**}(v^{Eb}_{t+1}) = \int_{-\infty}^{\infty} EZ^{E}(t,\bullet \mid v^{Eb}_{t+1}(m), v^{Et+1}_{t+1}, v^{N}_{t+1}) \frac{1}{\sqrt{\pi}} \exp -(m)^2 dm. \tag{B.1}$$

Thus, letting $Z^*(v^{Eb}_{t+1}) = EZ^E(t,\bullet \mid v^{Eb}_{t+1}(m),\bullet)/\sqrt{\pi}$, the approximation of $Z^{**}(v^{Eb}_{t+1})$ is given by:

$$\sum_{i=1}^{p} a_i Z^*(v^{Eb}_{t+1}(m_i)) \tag{B.2}$$

There are two sources of error associated with equation (B.2). The first error is that unless $Z^*(\bullet)$ is a polynomial of degree less than 2p-1, the quadrature method used gives only an approximation. Letting $v^{Eb}_{t+1,i}$ represent $v^{Eb}_{t+1,i}(m_i)$ this source of error can be written as:

$$e_1 = Z^{**}(v^{Eb}_{t+1}) - \Sigma_{i=1}^{p} a_i Z^*(v^{Eb}_{t+1,i}). \tag{B.3}$$

The second source of error arises because for a particular value of $v^{Eb}_{t+1,i}$, $Z^*(v^{Eb}_{t+1,i})$ is interpolated using the two nearest surrounding error points, $v^{Eb}_{t+1,i}+\psi$ and $v^{Eb}_{t+1,i}-(\Delta-\psi)$, where ψ is the distance

between $v^{Eb}_{t+1,i}$ and the next largest error point and Δ is the total distance between the two error points. That is, $Z^*(v^{Eb}_{t+1,i})$ is approximated by:

(B.4) $$\xi Z^*(v^{Eb}_{t+1,i}+\psi)+(1-\xi)Z^*(v^{Eb}_{t+1,i}-(\Delta-\psi)]$$

where $\xi=\psi/\Delta$. The second source of error can be written as:

(B.5) $$e_2= [\Sigma_{i=1}^{p} \xi Z^*(v^{Eb}_{t+1,i}+\psi_1) + (1-\xi)Z^*(v^{Eb}_{t+1,i}-(\Delta-\psi)) -Z^*(v^{Eb}_{t+1,i})].$$

The following two theorems establish convergence.

Theorem 1:

If $Z^*(v^{Eb}_{t+1})$ is continuous for all $v^{Eb}_{t+1}\varepsilon(-\infty,\infty)$ then

$\lim_{p\to\infty} e_1=0$.

Proof: It needs to be shown that for all $\varepsilon>0$ $\exists$ M such that $\forall p>M$, $|Z^{**}(v^{Eb}_{t+1})-\Sigma_{i=1}^{p}a_iZ^*(v^{Eb}_{t+1,i})|<\varepsilon$.

Any continuous function can be approximated arbitrarily well by a polynomial of high enough order (Secrest and Stroud, 1966). Formally, $\forall$ $\delta>0$ $\exists$ a polynomial, G_M, of degree M such that $\forall$ $v^{Eb}_{t+1,i}$ ε $(-\infty,\infty)$, $|Z^*(v^{Eb}_{t+1,i})-G_M(v^{Eb}_{t+1,i})|<\delta$.

By the triangle inequality,

$$|Z^{**}(v^{Eb}_{t+1})-\Sigma_{i=1}^{p}a_iZ^*(v^{Eb}_{t+1,i})|\leq |\int Z^*(v^{Eb}_{t+1})\exp\text{-}(m)^2dm - \int G_M(v^{Eb}_{t+1})\exp\text{-}(m)^2dm|$$
$$+|\int G_M(v^{Eb}_{t+1})\exp\text{-}(m)^2dm - \Sigma_{i=1}^{p}a_iG_M(v^{Eb}_{t+1,i})|$$
$$+|\Sigma_{i=1}^{p}a_iG_M(v^{Eb}_{t+1,i}) - \Sigma_{i=1}^{p}a_iZ^*(v^{Eb}_{t+1,i})|.$$

For p such that $2p-1\geq M$

$|\int Z^*(v^{Eb}_{t+1})\exp\text{-}(m)^2dm - \int G_M(v^{Eb}_{t+1})\exp\text{-}(m)^2dm| < \delta\int\exp\text{-}(m)^2dm$,

$|\int G_M(v^{Eb}_{t+1})\exp\text{-}(m)^2dm - \Sigma_{i=1}^{p}a_iG_M(v^{Eb}_{t+1,i})|=0$ and

$|\Sigma_{i=1}^{p}a_iG_M(v^{Eb}_{t+1,i}) - \Sigma_{i=1}^{p}a_iZ^*(v^{Eb}_{t+1,i})| <\delta \Sigma_{i=1}^{p}a_i= \delta\int\exp\text{-}(m)^2dm$.

The second line follows from the fact that G_M is a polynomial of order M, and the third line is true because the quadrature approximation of an integral of a constant function is exact for $p \geq 1$.

Thus for such m, $|Z^{**}(v^{Eb}_{t+1}) - \Sigma_{i=1}^{p} a_i Z^{*}(v^{Eb}_{t+1,i})| < 2\delta \int \exp\text{-}(m)^2 dm$. Then, for any $\varepsilon > 0$ we need to choose a value of M which implies a value of δ small enough to ensure that $2\delta \int \exp\text{-}(m)^2 dm < \varepsilon$. This implies that we need to choose the value of M which ensures that $\delta < 2\varepsilon / (\int \exp\text{-}(m)^2 dm)$.

Theorem 2: $\lim_{\Delta \to 0} e_2 = 0$.

Proof: This follows from the fact that as $\Delta \to 0$ the distance, ψ, between values of v^{Eb}_{t+1} and the nearest error points converges to zero.

$$\lim_{\Delta \to 0} \Sigma_{i=1}^{p} [\xi Z^{*}(v^{Eb}_{t+1,i} + \psi) + (1-\xi) Z^{*}(v^{Eb}_{t+1,i} - (\Delta - \psi)) - Z^{*}(v^{Eb}_{t+1,i})]$$

$$= \Sigma_{i=1}^{p} [\xi Z^{*}(v^{Eb}_{t+1,i}) + (1-\xi) Z^{*}(v^{Eb}_{t+1,i}) - Z^{*}(v^{Eb}_{t+1,i})] = 0.$$

Appendix C The procedure used for simulation of alternative wage policies.

For each of the individuals in the data, the following procedure is followed to simulate their choices for a particular wage policy. Given a set of estimates and the value functions for the person from the model, the set of error terms S_1 is drawn from its estimated distribution for the first period that the person is observed.[69] Given these error draws, there remains no uncertainty in the decision process, and the choice $j\varepsilon\{E1,N,H\}$ that the person makes can be found by comparing the values of the three options. Conditional on the choice of j, the set of error terms in the next period, S^j_2 is drawn.[70] The values of the elements of the relevant choice set, $\bar{D}^j_2$, are then compared to determine the decision that the individual would make in this period. This process is repeated in successive periods until the last year in which the individual is observed is reached.

[69]In period one the person does not have a previous job to return to.

[70]If j=E1 the serial correlation between the teaching wage error in successive periods is taken into account when drawing the value of S^j_2.

Appendix D Additional serially correlated variables

The non–parametric approach of approximating value functions using surrounding points remains applicable if more than one serially correlated, continuous error term exists. As before, when computing value functions by backwards recursion for the elements of the state grid at time t, it will be necessary to approximate value functions at time t+1 associated with points which are not in the t+1 state grid and, therefore, have not been solved by backwards recursion. What differs in this case is that the approximation must take place in multiple dimensions since the time t+1 value functions will be a function of the multiple error terms. Therefore, what is necessary in this case is to generalize the interpolating function to more than one dimension. One possible solution that easily generalizes to multiple dimensions is to use a local regression approach. As before, zero weight is assigned to the non–approximated points which are not "close" to the point for which a value function is being approximated. The effects of non–approximated points which "surround" (are "close" to) the point for which value functions are being interpolated are determined by a local, ordinary least squares regression.

The approximation algorithm used here can be controlled by the econometrician so that the subset of points which are chosen to be solved by backwards recursion can potentially depend on a person's observable years of experience and wages and an approximation level which is set by the researcher. The former ensures that value functions which are likely to have more of a direct effect on the likelihood function are calculated with more precision than other value functions.[71] This is potentially important in this application because many theoretically possible realizations of wage errors emerge from paths which are extremely unlikely to occur. In the

[71]That is, the "surrounding" points will be closer to the point which is being approximated.

spirit of Keane and Wolpin (1994) and Rust (1997), the latter implies that the econometrician has complete control over the number of state points for which value functions need to be solved by backwards recursion.[72]

Given the specification in this application, only one serially correlated state variable exists. However, the local regression approach applies to any state variable for which the econometrician wishes to approximate the value functions associated with some subset of its possible realizations at time t. Thus, time savings can be achieved and the method can be illustrated in this application by specifying the state grid to include combinations of some set of error points and a subset of the possible values of Y_t. Substantial time savings occur because interpolating the value functions associated with values of Y_t that are not in the state space is substantially faster than computing these value functions by traditional methods as part of the backwards recursion method. When error points are spaced an average of .4, and .80 of all possible years of experience values are included in the state grid, the average parameter estimate differs from the p=3 and Δ=.05 case by .16 of its standard error. When .70 of all possible years of experience values are included and error points are spaced an average of .4, the approximation bias statistic is .29. It is important to note that what appears to be useful is the general non–parametric "surrounding point" approach which is taken in this paper. There are other possible interpolating functions in multiple dimensions which would perhaps be be more consistent with the approach taken in one dimension in this paper. For example, Brien, Lillard, and Stern (1997) have recently proposed a weighting algorithm to deal with the interpolation of

[72]For any number of state variables, the econometrician could choose a precision level which implies that the number of value functions which need to be computed by traditional backwards recursion is arbitrarily small. Of course, one must be very concerned about the quality of the resulting approximation.

two continuous variables that does not involve an ordinary least squares regression, and, in some sense, is more similar in spirit to the interpolating approach which is taken in this paper when one serially correlated unobservable exists.

A More-Detailed Description of Value Function Interpolation in this Appendix

At each time t, the first step of the approximation involves selecting the subset of the state grid for which value functions will be calculated by backwards recursion. Suppose that the state grid, R_t, has been determined for all t. Using the same notation as in Appendix A, let R^Y_t be the set of possible years of teaching experience at time t, R^v_t be the set of error points at time t, $r^Y_t \varepsilon R^Y_t$, and $r^v_t \varepsilon R^v_t$. The interpolation algorithm determines a subset I^v_t of R^v_t and a subset I^Y_t of R^Y_t such that teaching value functions will be interpolated by OLS for all elements of the subset of R_t given by the Cartesian product $I^v_t \times I^Y_t$. The elements of the subset of R_t given by $\bar{I}^v_t \times \bar{I}^Y_t$, where $\bar{I}^v_t$ is the complement of I^v_t and $\bar{I}^Y_t$ is the complement of I^Y_t, are calculated by backwards recursion and are used in the interpolation of the elements of $I^v_t \times I^Y_t$. The algorithm ensures that grid points to be interpolated satisfy the following necessary condition which holds for j=v or Y.

condition (C.1) For any $r^j_t \varepsilon I^j_t$ there must exist elements r^{j*}_t and r^{j**}_t of $\bar{I}^j_t$ such that $r^j_t - r^{j*}_t < P_j(r^j_t)$ and $r^{j**}_t - r^j_t < P_j(r^j_t)$ where $P_j(\bullet)$ is a nonnegative "precision" function for j specified by the econometrician.[73]

[73]Condition (C.1) says that the precision function, $P_j(r^j_t)$, gives the maximum distance between a value of r^j_t which will be interpolated and the next largest and next smallest values that will be solved by backwards recursion and used in the interpolation of the grid point involving r^j_t. Thus, by decreasing the value of $P_j(r^j_t)$ the econometrician can increase the precision the of approximation.

The interpolation of the point $(r^Y_t, r^v_t) \in \Gamma^v_t \times \Gamma^Y_t$ first involves identifying the elements r^{Y*}_t, r^{Y**}_t, r^{v*}_t, and r^{v**}_t which, given r^Y_t and r^v_t, satisfy condition (B.1). OLS is then used to regress $\bar{V}^E(\bullet)$ on teaching experience and the teaching wage error using the observations (r^{Y*}_t, r^{v*}_t), (r^{Y*}_t, r^{v**}_t), (r^{Y**}_t, r^{v*}_t), and (r^{Y**}_t, r^{v**}_t) which are elements of $\bar{\Gamma}^v_t \times \bar{\Gamma}^Y_t$. Let α_0, α_1, and α_2 represent the estimated constant, the estimated coefficient on the years of teaching experience, and the estimated coefficient on the wage error term respectively in this regression. The interpolated teaching value function is then given by:

$$\bar{V}^E(t, \bullet, r^Y_t, r^v_t) = \alpha_0 + \alpha_1 r^Y_t + \alpha_2 r^v_t. \tag{C.2}$$

For an individual at time t, letting o^Y_t represent the observed value of Y_t and o^v_t represent the observed value of v^{Eb}_t, $P_j(r^j_t)$ is specified to be increasing in $|r^j_t - o^j_t|$ for j=v,Y. This gives the algorithm the desirable property that the closer (r^v_t, r^Y_t) is to (o^Y_t, o^v_t) the more precise the approximation will be. This is true because the OLS regression will involve points (r^{v*}_t, r^{Y*}_t), (r^{v*}_t, r^{Y**}_t), (r^{v**}_t, r^{Y*}_t), and (r^{v**}_t, r^{Y**}_t) which are closer to (r^v, r^Y).

References

Ballou, D. (1996): "Do Public Schools Hire the Best Applicants?" *Quarterly Journal of Economics*, pp. 97 133

Ballou, D. and Podgursky, M. (1995): "Recruiting Smater Teachers," *Journal of Human Resources*, 30, pp. 326-338.

Bellman, Richard, Dynamic Programming (Princeton: Princeton University Press, 1957)

Berkovec, James and Stern, Steven (1991): "Job Exit Behavior of Older Men," *Econometrica*, 59, 189-210.

Bendt, Ernst, Hall Bronwyn, Hall, Robert, and Hausman, Jerry (1974): "Estimation and Inference in Nonlinear Statistical Models," *Annals of Economic and Social Measurement.*

Bishop (1996): "Incentives to Study and the Organization of Secondary Instruction," published in Assessing Educational Practices, edited by William Becker and William Baumol, MIT press: Cambridge, Mass., Ch. 5.

Brewer (1996): "Career Paths and Quit Decisions: Evidence from Teaching," *Journal of Labor Economics*, 14, No. 2, 313-339.

Butler, J.S. and Moffitt, Robert (1982): "A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model," *Econometrica*, 50, 1982.

Cohn, E. (1996) Methods of Teacher Remuneration: Merit Pay and Career Ladders. Chapter 8: *Assessing Educational Practices: The Contribution of Economics*, edited by William E. Becker and William J. Baumol, MIT Press, Cambridge.

Brien, M., Lillard, L., and Stern, Steven (1997): "Cohabitation, Marriage, and Divorce in a Model of Match Quality," unpublished manuscript.

Cameron, S., and Heckman, J. (1998): "Lifecycles Schooling and Educational Selectivity: Models and Choice," *Journal of Political Economy*, April 1998, 108(2).

Dolton, P.J. (1990): "The Economics of UK Teacher Supply: The Graduate's Decision," *Economic Journal*, 100. pp. 91-104

Dolton, Peter and van der Klaauw, Wilbert (1995): "Leaving Teaching in the UK, A Duration Analysis," *The Economic Journal*, 105, 431-444.

Dolton, P.J. and Makepeace, G. (1993): "Female Labor Force Participation and the Choice of Occupation: The Supply of Teachers," *European Economic Review*, 37, pp.1393–1411.

Eberts, Randall (1987): "Union-negotiated Employment Rules and Teachers Quits." *Economics of Education Review*, Vol. 6 No. 1. pp. 15-25 1987

Engen, Eric (1992): *Precautionary Saving, Consumption, and Taxation in a Life Cycle Model with Stochastic Earnings and Mortality Risk.* Ph.D. Dissertation, University of Virginia.

Ehrenberg, P.J. and Makepeace, G.H. (1993): "Did Teacher's Race and Verbal Ability Matter in the 1960's," *Coleman*, Ithaca, NY: Cornell University, School of Industrial and Labor Relations: 1-57.

Erdem and Keane (1996): "Decision-making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets," *Marketing Science*, 15, 1, pp. 1-20.

Ferguson, R. (1990): "Racial Patterns in how School and Teacher Quality Affect Achievement and Earnings," Cambridge, Mass: Kennedy School of Government, Harvard University.

Giovannini, Alberto and Labadie, Pamela (1991): "Asset Prices and Interest Rates in Cash-in-Advance Models," *Journal of Political Economy*, 99, vol. 2, 1215-1251.

Hanushek, Eric A. (1971): "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro-Data," *American Economic Review*, 61(2): 280-288.

Hanushek, Eric A. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24, 1141-1177.

Heckman, J., and B. Singer, "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica* 52 (1984), 271–319.

Hodrick, Robert J., Kocherlakota, Narayana, and Lucas, Deborah (1991): "The Variability of Velocity in Cash-in-Advance Models," *Journal of Political Economy*, 99, no. 2, 358-384.

Hotz, Joseph V. and Miller, Robert A. (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *The Review of Economic Studies*, 60, 497-529.

Hotz, Joseph V., Miller, Robert A., Sanders, Seth, and Smith, Jeffrey. (1994): "A Simulation Estimator for Dynamic Models of Discrete Choice," *The Review of Economic Studies*, 61, 265-289.

Keane, Michael P. and Wolpin, Kenneth I. (1994): "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence," *Review of Economics and Statistics*, 76 648-672.

Lucas, R.E. Jr. (1976): "Econometric Policy Evaluation: A Critique," in K. Brunner and A.K. Meltzer. eds. *The Phillips Curve and Labour Markets*, Carnegie-Rochester Conference on Public Policy, North-Holland: Amsterdam.

Lumsdaine, R., Stock, J., and Wise D. (1991): "Three Models of Retirement: Computational Complexity Versus Predictive Validty." In D. Wise (ed.), Topics in the Economics of Aging, pp. 19-60. Chicago: University of Chicago Press.

Heckman, J. and Singer, B, "A Method for Minimizing the Impact of Distributional Assumption in Econometric Models for Duration Data," *Econometrica* 52(84), 217-319.

Hubbard, G., Skinner, J, and Zeldes, S. (1993): "Precationary Saving and Social Insurance," Journal of Political Economy, 1995, vol. 103, no. 2.

MaCurdy, Thomas E. "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis," Journal of Econometrics, 18: 83-114.

Manski, Charles F. and McFadden, Daniel (1981): *Structural Analysis of Discrete Data with Econometric Applications*, (Cambridge Mass.: The MIT Press, 1981).

Miller, Robert, "Job Matching and Occupational Choice," Journal of Political Economy 92 (Dec. 1984), 1086-1120.

Monk, D. (1992): "Subject Area Preparation of Secondary Mathematics and Science Teachers and Students Achievement," Department of Education, Cornell University: 1-51.

Mont, Daniel and Rees, Daniel (1996): "The Influence of Classroom Characteristics on High School Teacher Turnover," *Economic Inquiry*, 34, 152-167.

Murnane, Richard J. and Olsen, Randall J. (1989): "The Effects of Salaries and Opportunity Costs on Duration in Teaching: Evidence from Michigan," Review of Economics and Statistics, 71(2), 347-352.

Murnane, Richard J. and Olsen, Randall J. (1990): "The Effects of Salaries and Opportunity Costs on Length of Stay in Teaching, Evidence from North Carolina," *The Journal of Human Resources*, 25, 106-124.

Murnane, Richard J., Singer, Judith D., and Willett, John B. (1989b) "The Influences of Salaries and Opportunity Costs on Teachers' Career Choices: Evidence from North Carolina," *Harvard Educational Review*, 59, 345-346

Murnane, Richard J., Singer, Judith D., Willett, John B., Kemple, James J., and Olsen, Randall J., *Who Will Teach? Policies That Matter* (London: Harvard University Press, 1991).

Murnane, Richard J., Willett, John B., and Levy, Frank (1995) "The Growing Importance of Cognitive Skills in Wage Determination," NBER Working Paper, no. 5076.

Mroz, Tom and Guilkey, David (1992) "Discrete Factor Approximations for Use in Simultaneous Equation Models with Both Continuous and Discrete Endogenous Variable," mimeo

Pakes, Ariel (1987): "Patents as Options: Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 54, 755-784

Palumbo, Michael (1991): *Health Uncertainty and Optimal Consumption Near the End of the Life Cycle*. Ph.D. Dissertation, University of Virginia.

Rust, John (1987): "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 55, 999-1033.

Rust , John (1992): "Dynamic Structural Models: Problems and Prospects, Part I: Discrete Decision Processes." In Laffont, ed., *Advances in Econnometrics, Proceedings of the 6th World Congress of the Econometric Society*. New York: Cambridge University Press, 1992.

Rust, John (1997): "Using Randomization to Break the Curse of Dimensionality," Econometrica, Voll. 65, No 3, May 1997, pp. 781-832.

Rust, John and Phelan, Christopher (1997): "How Social Security and Medicare affect retirement behavior in a World of Incomplete Markets," *Econometrica*, Vol 65, No. 3, May 1997, pp. 487-516.

Sickles, Robin C. and Taubman, Paul (1986): "An analysis of the Health and Retirement Status of the Elderly," *Econometrica*, 54, 1339-1356.

Stern, Steven (1997): "Approximate Solutions to Stochastic Dynamic Programs." Econometric Theory, June 1997.

Stern, Steven (1997b): "Simulation Based Estimation," Journal of Economic Literature, December, 1997.

Stinebrickner, Todd R. (1998): "An Empirical Investigation of Teacher Attrition," forthcoming, *Economics of Education Review*.

Stinebrickner, Todd R. (1996a): "A Dynamic Model of Teacher Labor Supply," unpublished manuscript.

Stinebrickner, Todd R. (1996b): "An Analysis of Occupational Change and Departure from the Labor Force: Evidence of the Reasons that Teachers Leave," unpublished manuscript.

Stinebrickner, Todd R. (1996c): "Estimating an Endogenous Index to Deal with Vectors of Characteristics in a Dynamic, Discrete Choice Model: An Application Involving the Characteristics of U.S. Schools," unpublished manuscript.

Strauss, Robert P. and Sawyer, Elizabeth A. (1986): "Some New Evidence on Teacher and Student Competencies," *Economics of Education Review*, 5(1): 41-48.

Stock, J. and Wise, D. (1990): "Pensions, the Option Value of Work, and Retirement, Econometrica, 58, 1151-1180

Stroud, A.H. and Secrest (1966): *Gaussian Quadrature Formulas*, Prentice Hall, Englewood Cliffs, NJ.

Swann, Chris (1996): "A Dynamic Analysis of Marriage, Labor Force, Work, and Participation in AFDC," mimeo.

Tauchen, George and Hussey, Robert (1991): "Quadrature-Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models," *Econometrica*, 59: 371-396.

Theobald, Neil D. (1990): "An Examination of the Influence of Personal, Professional, and School District Characteristics on Public School Teacher Retention," *Economics of Education Review*, 9, 241-250.

Theobald, N.D. and Gritz, M.R. (1996): The Effects of School District Spending Priorities on the Exit Paths of Beginning Teachers Leaving the District. *Economics of Education Review* 15(1), 11-22.

Van der Klaauw, W. (1996a): "Expectations and Career Decisions: An Analysis of Teaching Careers Using Expectations Data," mimeo.

Van der Klaauw, W. (1996b): "Female Labor Supply and Marital Status Decisions: A Life Cycle Model," Review of Economic Studies, 63, No. 215, pp. 199-236

figure 1

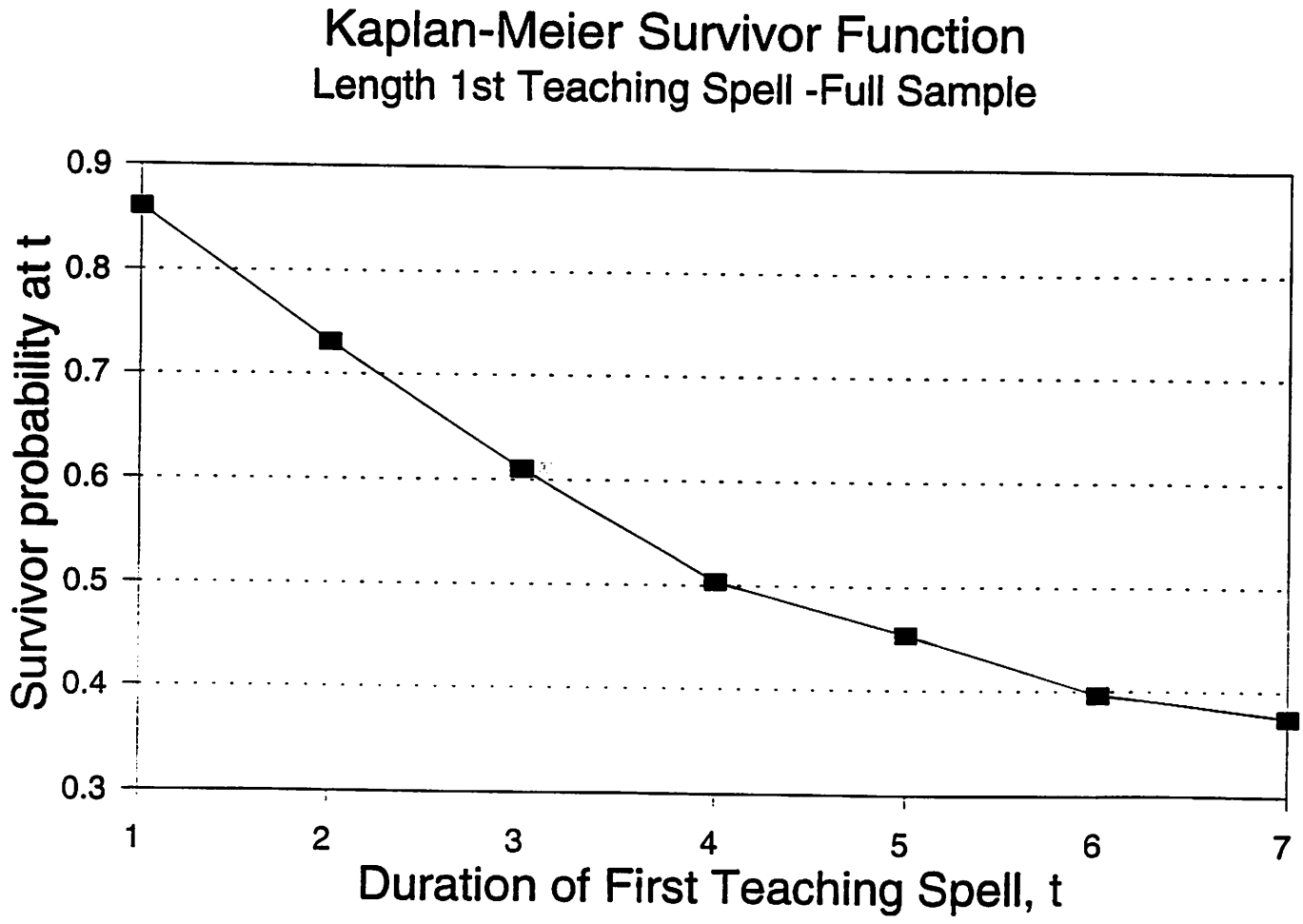

Kaplan-Meier Survivor Function
Length 1st Teaching Spell -Full Sample
Survivor probability at t
0.9
0.8
0.7
0.6
0.5
0.4
0.3
1
2
3
4
5
6
7
Duration of First Teaching Spell, t

figure 2

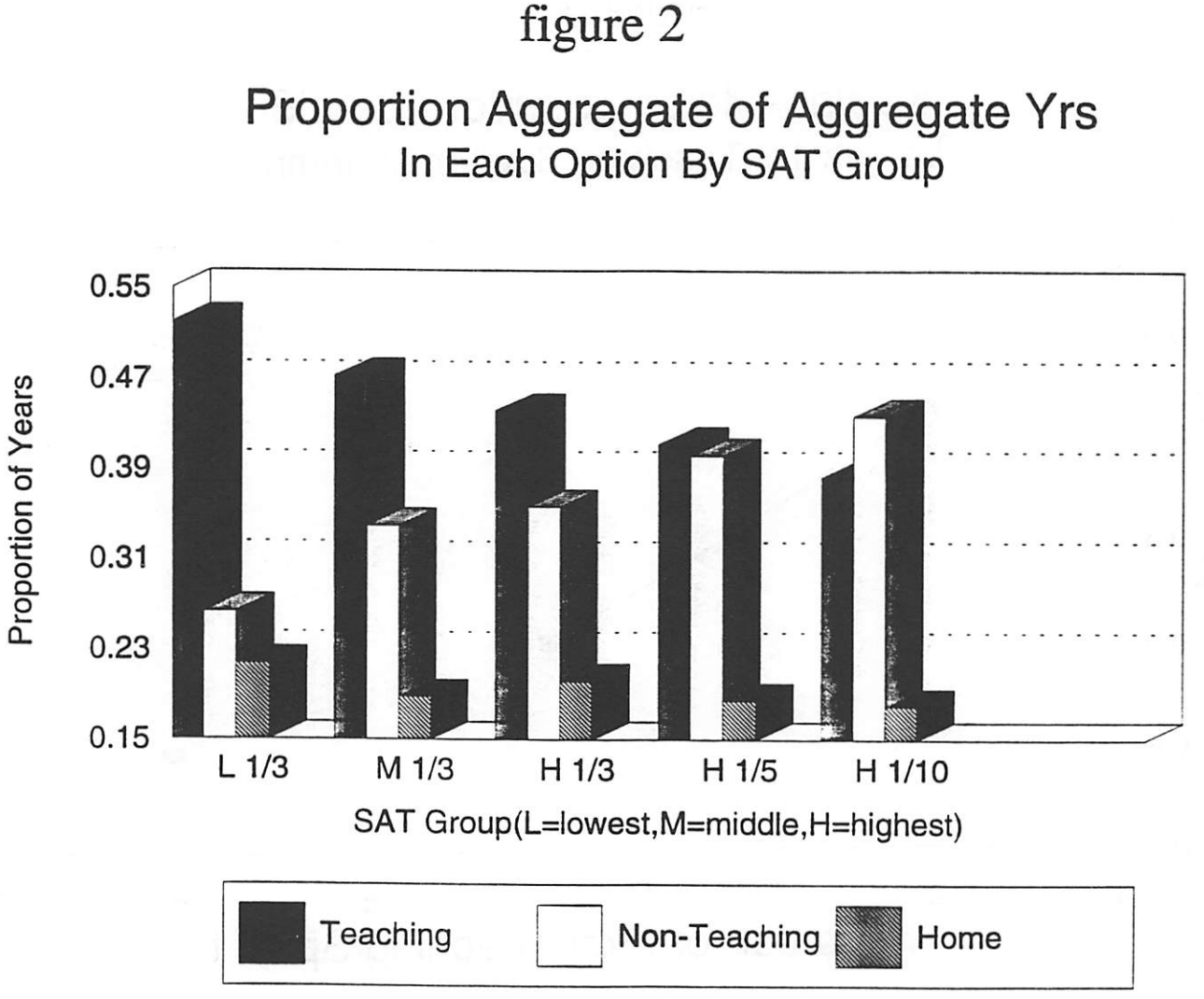

Proportion Aggregate of Aggregate Yrs
In Each Option By SAT Group
Proportion of Years
0.55
0.47
0.39
0.31
0.23
0.15
L 1/3
M 1/3
H 1/3
H 1/5
H 1/10
SAT Group(L=lowest,M=middle,H=highest)
Teaching
Non-Teaching
Home

figure 3

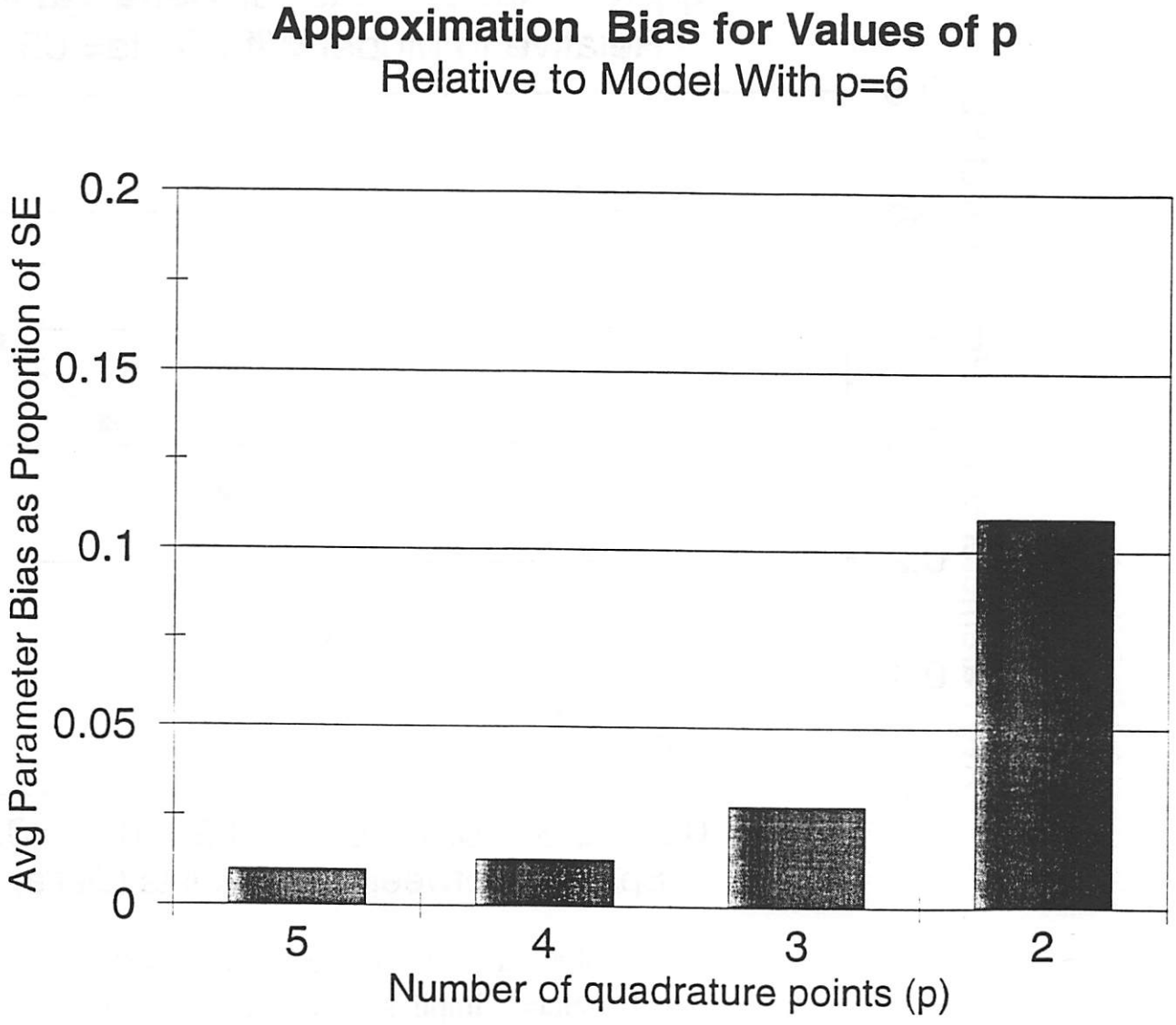
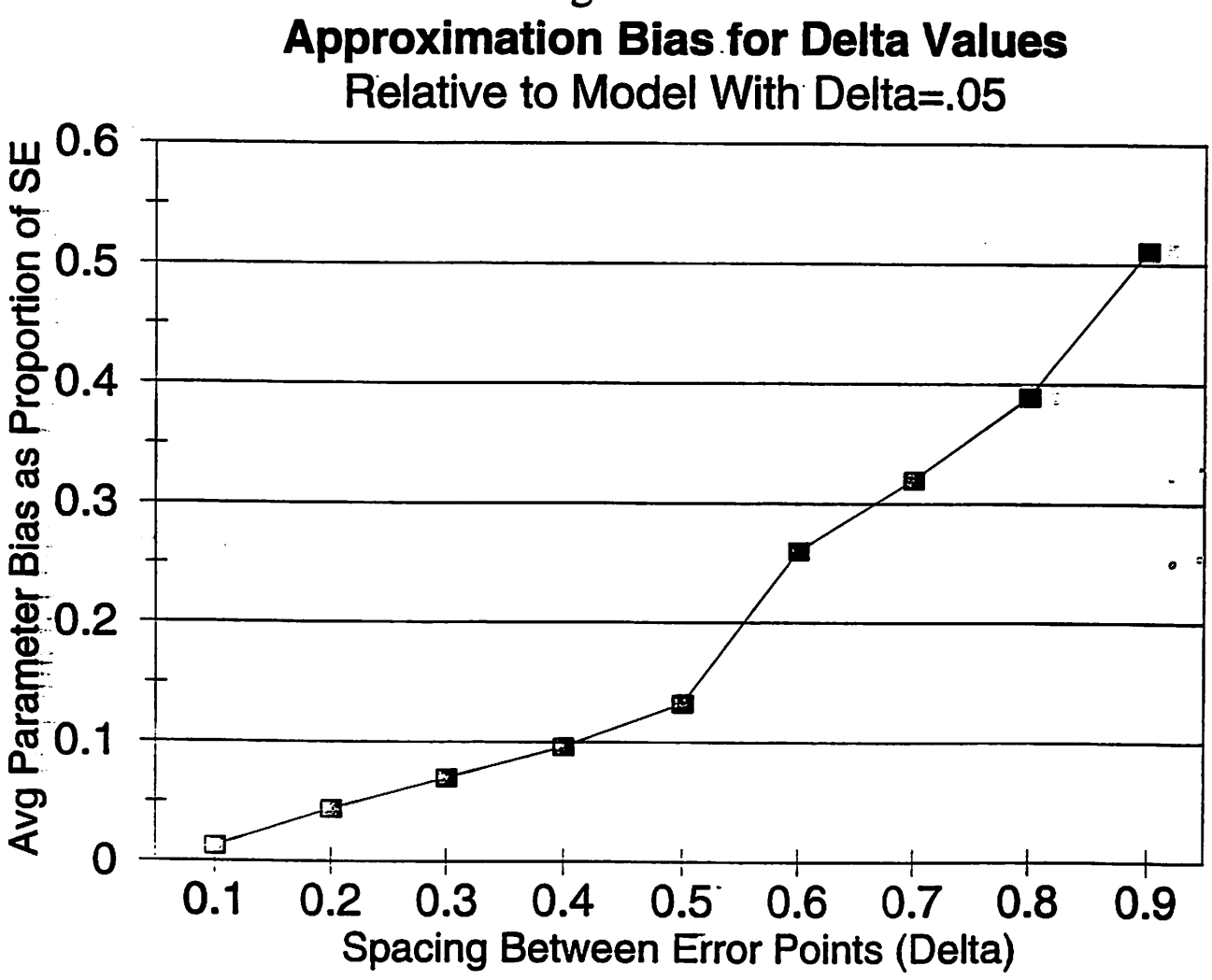

Approximation Bias for Values of p
Relative to Model With p=6
Avg Parameter Bias as Proportion of SE
0.2
0.15
0.1
0.05
0
5
4
3
2
Number of quadrature points (p)

figure 4



note: the standard deviation of wage =.35

e.g., delta=.4 implies points are 1.14 s.d.'s apart

figure 5

Proportion Aggregate Years Each Option Full-Sample for Different Policies

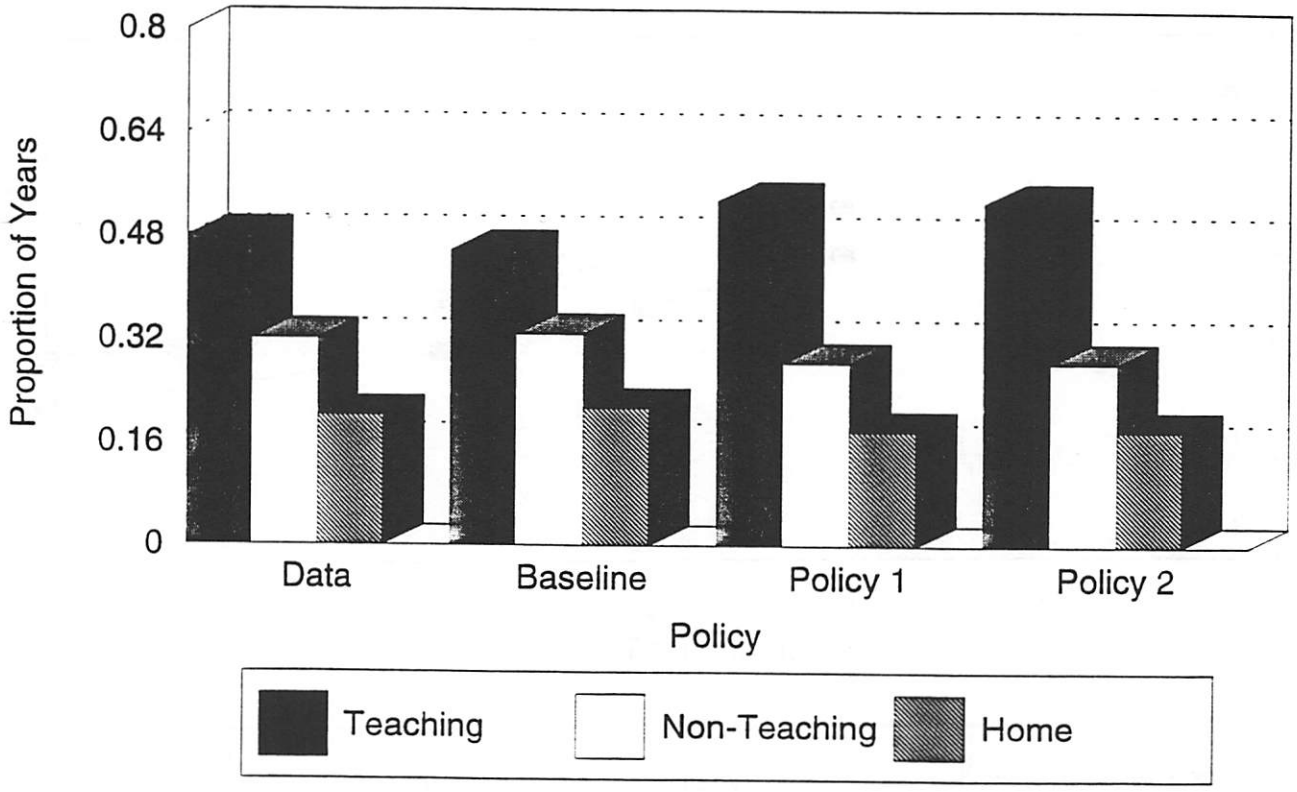


figure 6

Survivor Functions for Full Sample

For Different Policy Simulations

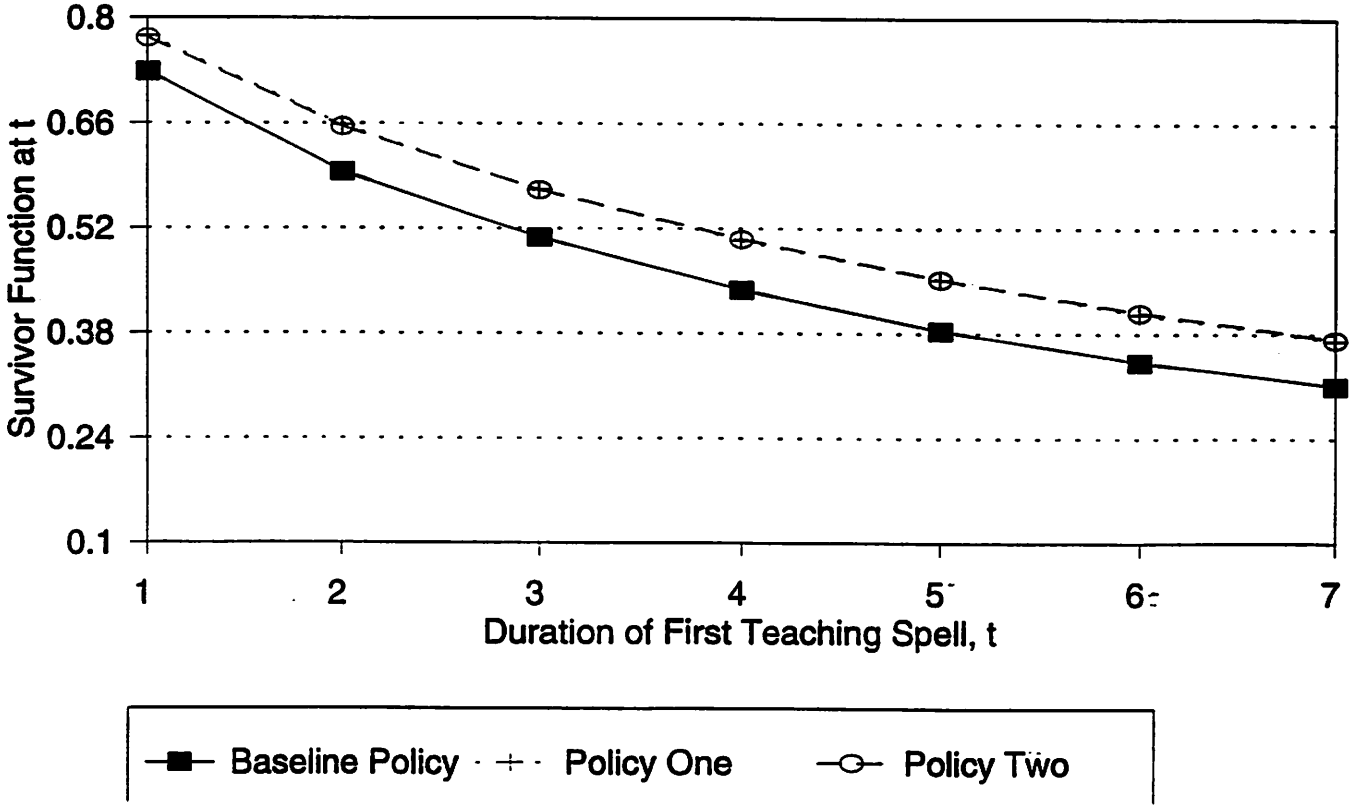


figure 7

Relative Labor Supply of High SAT Grou

Figure shows proportion of aggregate years that individuals in low SAT group choose a particular option divided by the proportion of aggregate years that individuals in high SAT group choose the option.

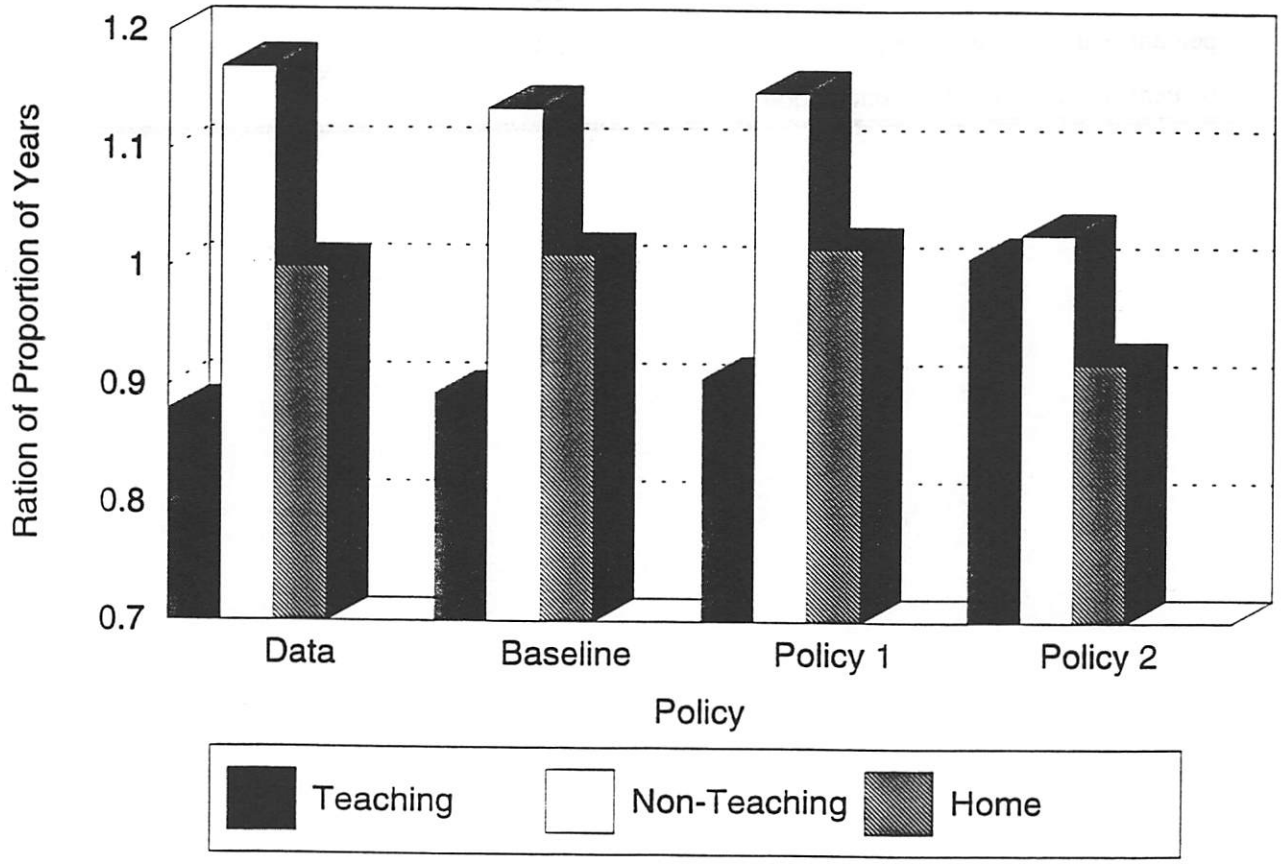


TABLE 1
DESCRIPTIVE DEMOGRAPHIC STATISTICS

variable	mean	standard deviation
number of years individual is observed (after certification)	9.0	4.1
math SAT	476	93
number of children (as of 1986)	1.1	1.1
number of years of post-bachelor education (as of 1986)	1.05	1.2
percent female	72.5	
percent married in a given period	63.4	
percent married in at least one period	81.4	

Table 2 RANGE OF POSSIBLE ERROR REALIZATIONS IN EACH PERIOD for Hypothetical Values of ρ

	ρ=.91		ρ=.70		ρ=.50	
Year (2=1976)	**min**	**max**	**min**	**max**	**min**	**max**
2	-1.12	1.13	-1.11	1.12	-1.10	1.10
3	-1.45	1.46	-1.30	1.31	-1.10	1.10
4	-1.82	1.82	-1.50	1.50	-1.10	1.10
5	-2.15	2.15	-1.61	1.62	-1.10	1.10
6	-2.52	2.53	-1.71	1.72	-1.10	1.10
7	-2.82	2.83	-1.75	1.76	-1.10	1.10
8	-3.13	3.13	-1.77	1.77	-1.10	1.10
9	-3.42	3.43	-1.78	1.78	-1.10	1.10
10	-3.70	3.71	-1.78	1.79	-1.12	1.12
11	-3.92	3.92	-1.78	1.78	-1.12	1.12
12	-4.14	4.15	-1.79	1.79	-1.12	1.12
13	-4.36	4.36	-1.79	1.79	-1.12	1.12
14	-4.57	4.57	-1.80	1.80	-1.12	1.12
15	-4.78	4.78	-1.80	1.80	-1.12	1.12
16	-4.98	4.98	-1.80	1.80	-1.12	1.12
17	-5.13	5.13	-1.80	1.80	-1.12	1.12
18	-5.25	5.25	-1.80	1.80	-1.12	1.12

Table shows average range of possible error realizations that could occur in each period for hypothetical values of ρ. This range determines how many error points exist for a particular choice of spacing between the points, Δ.

Table 3 COMPARISON OF P=6, Δ=.1 AND P=3, AVGΔ=.4 ESTIMATES
model without heterogeneity and without interaction terms

Variable	Parameter Estimates p=6 Δ=.1	Parameter Estimates p=3 avgΔ=.4	Approximation Bias	Estimated Standard Error of Parameter	Bias/SE
Teaching Wage					
CONST	5.42402	5.42908	.00505	.04431	.11412
TIME	-.16343	-.16545	.00201	.01237	.16309
TIMESQ	.00643	.00565	.00013	.00092	.14330
MALE	.00470	.00493	.00023	.01237	.01881
SAT	-.01037	-.01020	.00016	.00660	.02528
EDU	.03698	.03690	.00008	.00757	.01089
EXP	.02464	.02504	.00039	.00442	.08997
ρ	.91371	.91370	.00001	.00914	.00135
Non-teaching WAge					
CONST	4.40347	4.40394	.00046	.09219	.00508
TIME	-.01740	-.017291	.00011	.02622	.00441
TIMESQ	.00565	.00564	.00001	.00188	.00619
MALE	.09059	.09052	.00007	.02754	.00278
SAT	.03964	.03968	.00001	.01187	.00112
EDU	.05674	.056697	.00005	.00879	.00579
Teaching Non-Wage					
CONST	-4.56980	-4.57734	.00753	.08232	.09159
MALE	.46312	.46492	.00180	.03549	.05073
SAT	.023075	.02301	.00006	.01339	.00491
EXP	-.00397	-.00361	.00036	.00616	.05919
CHILD	-.20371	-.20386	.00015	.01287	.01194
MARR	-.26375	-.26510	.00134	.03223	.04180
Non-Teaching Non-Wage					
CONST	-4.3567	-4.35604	.00071	.08362	.00852
MALE	.45009	.45202	.00192	.04130	.04654
SAT	.01462	.01482	.00020	.01596	.01264
EXP	-.13448	-.13475	.00027	.00654	.04216
CHILD	-.20021	-.20084	.00053	.01403	.03793
MARR	-.31423	-.31504	.00117	.03460	.03381
Var Terms					
σ_E	.35118	.35113	.00004	.00392	.01231
σ_e	.30925	.30786	.00138	.00455	.30329
σ_n	.46008	.46034	.00025	.00670	.03826
τ	.42123	.42324	.00214	.012071	.166867
					Avg (Bias/SE)
					.05182

TABLE 4 STRUCTURAL MODEL ESTIMATES -
model with unobserved heterogeneity (p=3, avgΔ=.4)

VARIABLE	
Teaching Wage	
CONST	5.765* (.045)
TIME (1975=1, 1976=2,...)	-.272* (.013)
TIME*TIME	.015* (.001)
MALE	.014 (.012)
SAT	-.012 (.007)
EDU	.060* (.009)
EXP	.016* (.005)
Autoregressive Coefficient	
ρ	.924* (.007)
Non-Teaching Wage	
CONST	4.446* (.097)
TIME (1975=1,1976=2....)	-.003 (.027)
TIME*TIME	.004* (.002)
MALE	.091* (.029)
SAT	.041* (.012)
EDU	.030* (.014)

Teaching Non-Pecuniary Utility	
CONST	-3.700* (.178)
MALE	.101 (.100)
SAT	-.083* (.028)
EXP	-.048* (.008)
CHILD	-.422* (.024)
CHILDXMALE	.377* (.063)
MARR	-.408 (.069)*
MARRXMALE	.300* (.111)
Non-Teaching Non-Pecuniary Utility	
CONST	-4.127* (.161)
MALE	-.267* (.107)
SAT	-.043 (.030)
EXP	-.059* (.007)
CHILD	-.458* (.026)
CHILDXMALE	.487* (.060)
MARR	-.373* (.069)
MARRXMALE	.653* (.119)

Variance Terms

σ_1	(heterogeneity teaching)	.406*
		(.034)
σ_2	(heterogeneity non-teaching)	.778*
		(.044)
σ_3	(heterogeneity home)	.565*
		(.035)
σ_E		.347*
		(.004)
σ_e		.307*
		(.004)
σ_N		.460*
		(.007)
τ		.417*
		(.006)
Log Likelihood Function		-4345.37

The numbers are estimates from a specification in which the discount factor, β, is set to .95 The numbers in parentheses are asymptotic standard errors. * denotes an asymptotic t ratio greater than two.